A Cluster-based Model of COVID-19 Transmission Dynamics

B SHAYAK (*)

Theoretical and Applied Mechanics,
Mechanical and Aerospace Engg,
Cornell University,
Ithaca – 14853,
New York State, USA

MOHIT M SHARMA

Population Health Sciences,
Weill Cornell Medicine,
1300 York Avenue,
NYC – 10065,
New York State, USA

(*) Corresponding author. Email : sb2344@cornell.edu , shayak.2015@iitkalumni.org
ORCID : 0000-0003-2502-2268

Thursday 03 June 2021
Current revision Tuesday 05 October 2021

---- o ----

Keywords. Interaction heterogeneity Deterministic model Parameter sensitivity Initial condition sensitivity

Abstract. Many countries have manifested COVID-19 trajectories where extended periods of constant and low daily case rate suddenly transition to epidemic waves of considerable severity with no correspondingly drastic relaxation in preventive measures. Such solutions are outside the scope of classical epidemiological models. Here we construct a deterministic, discrete-time, discrete-population mathematical model called *cluster seeding and transmission (CST) model* which can explain these non-classical phenomena. Our key hypothesis is that with partial preventive measures in place, viral transmission occurs primarily within small, closed groups of family members and friends, which we label as *clusters*. Inter-cluster transmission is infrequent compared to intra-cluster transmission but it is the key to determining the course of the epidemic. If inter-cluster transmission is low enough, we see stable plateau solutions. Above a cutoff level however, such transmission can destabilize a plateau into a huge wave even though its contribution to the population-averaged spreading rate still remains small. We call this the *cryptogenic instability*. We also find that stochastic effects when case counts are very low may result in a temporary and artificial suppression of an instability; we call this the *critical mass effect*. Both these phenomena are absent from conventional infectious disease models and militate against the successful management of the epidemic.

Word counts. Abstract – 207, Main Text – 12381, Grand total – 17202

---- o ---- o ---- o ----   ---- o ---- o ---- o ----

Multiple waves have characterized the COVID-19 epidemic trajectories in many countries. USA has had five waves of which the third has been the most severe so far. Most European nations had their second waves in mid-to-late 2020 and in many cases these were more severe than the respective first waves. In April-May 2021, after months of very low case counts, India exhibited a massive second wave which dwarfed the first one in terms of size and rapidity and led to tens or even hundreds of thousands of avoidable deaths. In May 2021, Taiwan began showing its first wave where there were more fatalities every day than there had been over the past fifteen months combined. In many cases, the onset of the epidemic waves is sudden, does not follow any relaxation of non-pharmaceutical interventions and appears to defy logic. We propose a model, centred on heterogeneity in human interaction patterns, which explains the sudden origins of epidemic waves from a nearly quiescent background. We present the city of Delhi, India as a case study for this model.

INTRODUCTION

§1. Introduction to COVID-19 mathematical models. A huge literature exists on the mathematical modeling of COVID-19 dynamics. Below we give the briefest possible introduction to this literature, highlighting the studies which have exerted considerable influence and/or reported significantly novel content. The models in use can be classified into several broad types :

(1) Compartmental or lumped parameter models : These use ordinary, delayed or in some cases partial differential equations to describe the spread of the disease. The first ever infectious disease dynamic model, the S-I-R model [1], was of this type. This model and its variant the S-E-I-R model have been used extensively to describe and forecast COVID-19 trajectories, for example in Ref. [2]. These models suffer from an apparent flaw in that many of the parameters are not directly related to the disease, its properties and interventions. For example, in the S-E-I-R model, there is no parameter which explicitly incorporates the asymptomatic infection period or a contact tracing campaign. To overcome this limitation, new compartments, for example for asymptomatic and hospitalized cases, are often added to the basic structure to form a complex equation system with many variables and parameters. References [3-8] are studies of this class which have attracted considerable attention. While these are more powerful than the basic models, they still have certain limitations.

Firstly, the structure $\mathrm{d}I/\mathrm{d}t = -\gamma I + (.....)$ with γ being a rate constant necessarily implies that, if there is an initial infected pool and zero fresh infections, then the infected population will decay exponentially in time. In reality, this will not be the case – if new infections are somehow ceased (hypothetically), then almost the entire infected population will recover after the infection period. In other words, infection states (susceptible, infected etc) are memoryless – the duration for which a person remains in a particular state is independent of how long s/he has already been in that state. This might constitute a difficulty for states such as “infected” and “quarantined”. Secondly, the calculation of the reproduction number R (the number of secondary cases derived from one primary case) for these ODE models at any point other than the starting one is not well documented.

To overcome these limitations, our group has developed a compartmental model based on delay differential equations (DDE) [9]. We believe that, among models of this class, DDE model is considerably realistic and has high descriptive and predictive power. The reasons for our this belief have been described in detail in our publication; taking these for granted, we shall in the rest of this work treat this model as a benchmark compartmental model, and representative of all models of this type. We shall also use the terminology “classical” to refer to compartmental models per se.

(2) Agent-based models : These models treat people as lattice sites on a network, and infect a person with certain probabilities if his/her neighbours are infected. These are by far the most detailed models and are capable of capturing details which no other model can account for. The level of granularity in these models can vary widely; among the more generic examples we have Refs. [10,11] while city- or country-specific models include Refs. [12-15]. The increased accuracy of these models comes with a few trade-offs

however. Firstly, the models are computationally intensive, requiring high-performance or cluster computers for operation. Secondly, the results are extremely sensitive to the network structure assumed by the modellers. Thirdly, city-specific formulations are often inapplicable to other regions and hence cast limited insight into the dynamics of the contagion in general.

(3) Other models : These include stochastic differential equation models [16,17] and data-driven models [18]. So far, they have not achieved any significant descriptive or predictive success. In addition, there are network models of generic epidemic propagation [19-23], written prior to the advent of the COVID-19 pandemic, which have so far not demonstrated practical relevance in the present situation.

§2. Delhi, India and the limitations of compartmental models. The COVID-19 wave in April 2021 in the city of Delhi, India acted as a tragic demonstration of the limitations of existing coronavirus models including Ref. [9]. During 2020, this city underwent two waves of the pandemic, while India as a whole exhibited only one wave. The case trajectory, i.e. number of new cases being reported per day, in Delhi between 01 August and 31 December 2020 is shown in Fig. 1. The first wave in this Figure was caused by a relaxation during the phasewise exit from lockdown which commenced in India beginning 08 June, while the second was related to festive season. The festival Dussehra was on 26 October and Diwali on 14 November, which correspond well to the second peak. For both waves, the rate of ascent was slow with about a threefold increase in cases over the span of a month, approximately corresponding to a reproduction number $R = 1.2$. Renewed emphasis on non-pharmaceutical interventions (NPI) resulted in a de-escalation in each case, well before the outbreak morphed into a public health crisis.

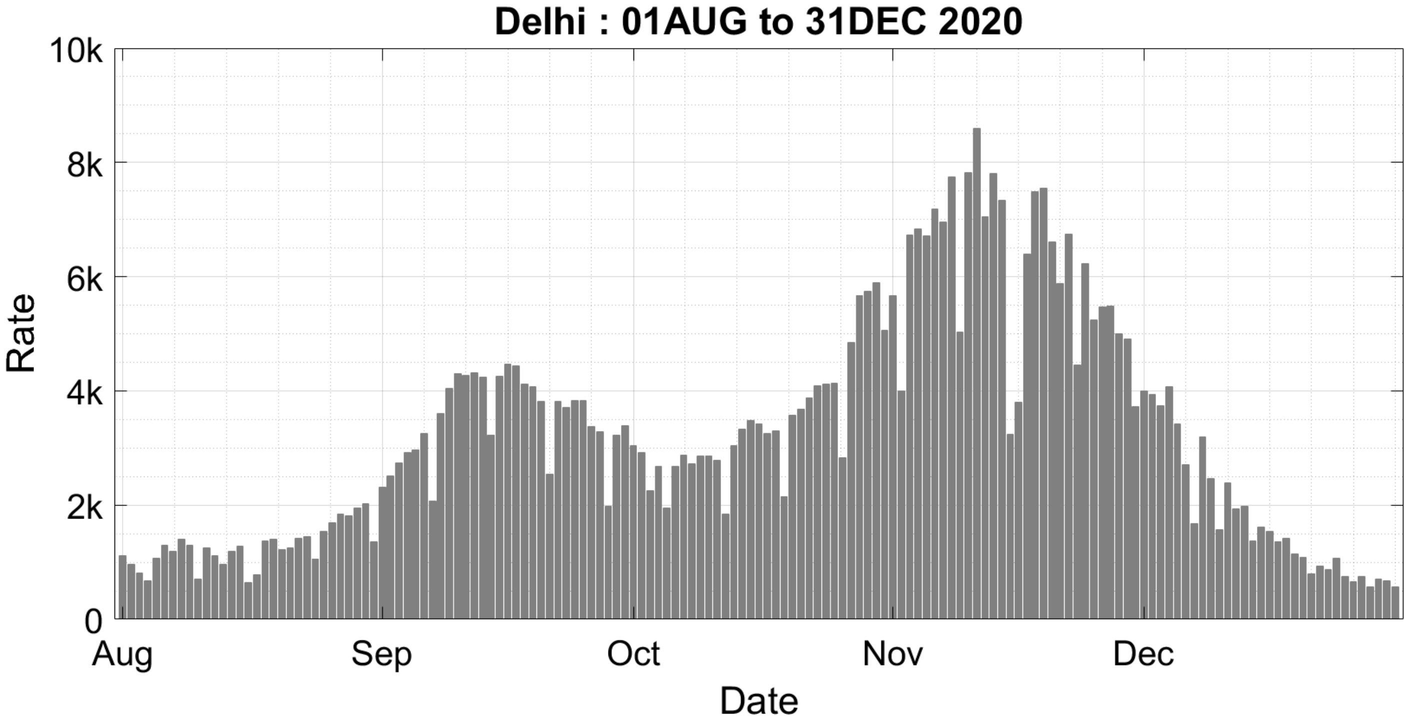

Figure 1 : *Case trajectories in Delhi during the period indicated. The symbol 'k' denotes thousand. Labels on the x-axis correspond to the first day of each month.*

By late January 2021, there were less than 200 daily cases in Delhi even though many NPI had been relaxed by then. Resumption of public transport services including Delhi Metro and reopening of restaurants, bars, cinema halls and the like had already taken place during November and December while cases decreased almost monotonically – soon after opening, entertainment venues were operating at or near capacity. The size of gatherings for wedding parties, funeral functions etc was raised and people started participating in these activities with enthusiasm. Mask and separation mandates had never been officially released but by January, compliance had waned significantly; despite this, cases remained very low and nearly constant all through February. Conventional epidemiological models could explain this phenomenon in only one way, which was through herd immunity. The scientific understanding ran that enough undetected cases had occurred in 2020 itself to bring the population of Delhi close to the herd immunity threshold, and that the risk of a new wave was now past. The successful management of the festival-related waves also gave rise to a false sense of confidence.

During February-March 2021, cases in the state of Maharashtra rose gradually but significantly, signalling the beginning of a second wave. Even so, there was no alarm sounded in Delhi or anywhere else in the country. This too was consistent with conventional epidemiological understanding – if Delhi was herd immune then an influx of cases from outside could not topple it from its secure position. In mid-March however, things started deteriorating and, unlike in Maharashtra, did so with extreme rapidity. From less than 400 daily cases on 10 March, Delhi climbed to 800 on 20 March, then to 2800 on 01 April, 8000 on 10 April and finally a record peak of 28,000 on 20 April. A full lockdown was declared 2-3 days before the peak but by then healthcare systems had already been strained beyond their limits. Figure 2 shows the epidemic trajectory in Delhi between 01 March and 30 April 2021.

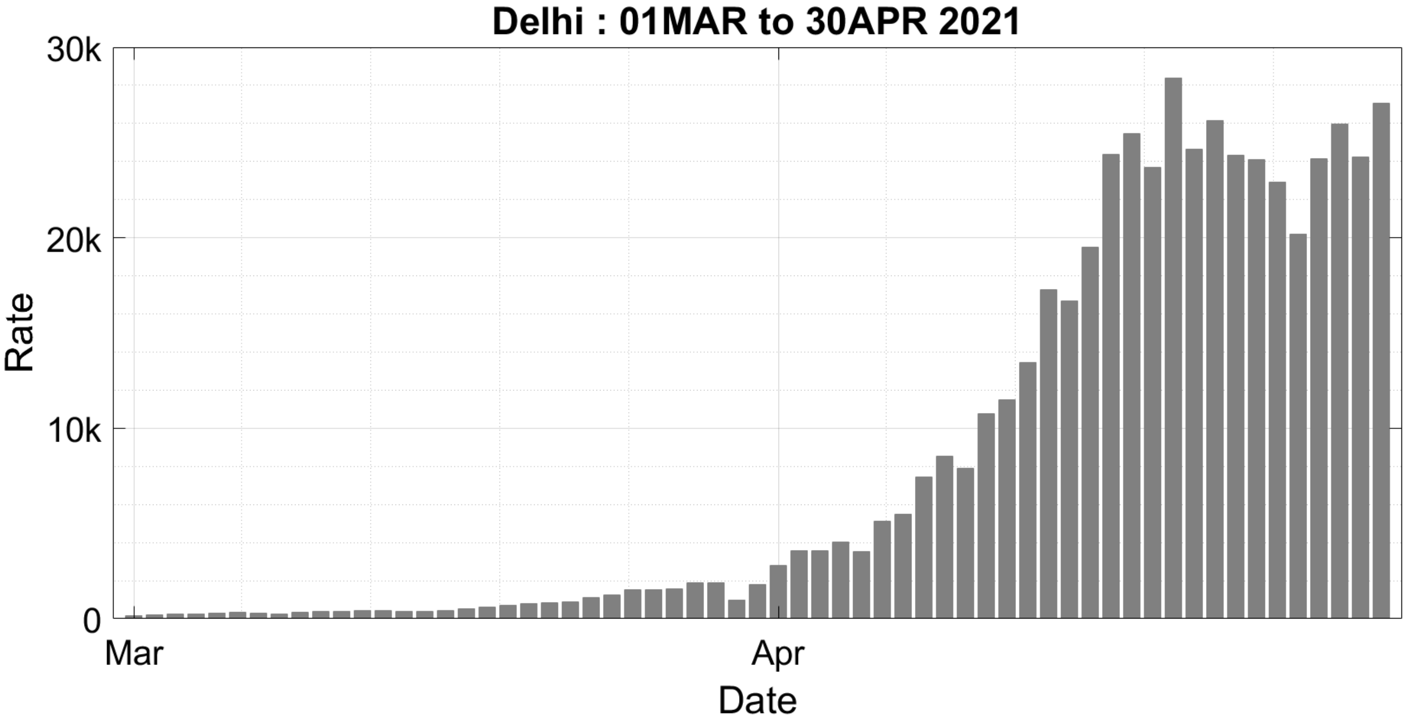

Figure 2 : *Case trajectories in Delhi during the period indicated. The symbol 'k' denotes thousand. Labels on the* x*-axis correspond to the first day of each month.*

Classical mathematical modeling of epidemics makes us expect an increase in case counts if there is a relaxation of NPI or a reduction in COVID-appropriate behaviour, and a decrease in case counts if there is a strengthening of NPI or a positive trend in compliance. The August and October waves in Delhi shown in Fig. 1 are of this type. The third wave in USA, fuelled by travelling and festivities during the Thanksgiving-Christmas-New Year season, was also in accordance with this understanding. Even though the quantitative predictions of many models for both India and USA were not sufficiently accurate, these waves did not indicate a conceptual gap. The reality of Delhi in early 2021 was however completely different. Decrease and plateauing of cases during phased relaxation of NPI and a massive wave with no change in NPI was fundamentally unexpected. Just as surprising was the suddenness of the wave – even when an increase in mobility causes R to cross unity, we expect a gradual increase in cases, at least at first. This increase indicates over-reopening and gives time for a rollback of relaxations before there is a public health crisis. Such an expectation lay at the core of the phased emergence from lockdown which was implemented in nearly every country worldwide, and was in line with the festival waves in Delhi. The April wave however, where the increase began at $R = 1.8$ from the first day itself, was completely new. Nothing even remotely similar was predicted by any existing mathematical model.

There was only one serious effort at an explanation of the Delhi wave, which was made by Dhar et. al. [24]. After a detailed analysis of genomic sequences, this study concluded that the double mutant B1.617.2 ("delta" variant) was responsible for the disaster. While there is no doubt that this mutant greatly exacerbated the disease burden, there are a few observations which suggest that it was not the entire story. These are as follows.

- Double mutant strain was first discovered in Maharashtra in October 2020. Case counts there and elsewhere decreased for a long time thereafter. Why did the mutant remain dormant for so many months before suddenly initiating uncontrollable spread ? Continuing on these lines, double mutant was introduced to Delhi from Maharashtra as per Dhar et. al. Why is the reproduction number in Delhi more than 50 percent higher than in Maharashtra, and especially in its two cities Mumbai and Pune which are socio-culturally very similar to Delhi ?

- As per Fig. 3A of Dhar et. al. [24], wildtype or B1 amounted to about 15 percent of total cases in March and about 10 percent of total cases in Delhi in April 2021. Given that there were at least 25 times more cases in April than in March, this implies that wildtype infections also underwent significant growth during the second wave, albeit at a slower rate than double mutant. This is impossible if the mutant is the only factor driving increased transmission.

Although Delhi (and many other cities in India) provided horrific examples on account of their high population, sudden waves have been manifest in other places as well. During writing of the original version of this Article, Japan was being hit by a fourth wave, placing the Olympics under threat. In May-June 2021, Taiwan exhibited its first huge wave of the disease after more than a year of staying at nearly zero level. In many instances, like Delhi, the waves emerged from a "plateau" or an extended period of nearly constant daily case rate. Plateaus are outside the ambit of lumped parameter models, including Ref. [9]. In these models, a constant case rate is not a generic solution; it occurs only if R equals exactly unity, which, for a given set of parameter values, happens for just one particular infection level. These phenomena indicated that a more sophisticated model, which could at least qualitatively capture the plateau and sudden wave solutions, was the need of the hour.

§**3. Models with plateaus.** A handful of models so far have succeeded in finding a plateau as a generic solution; we enumerate these in this Section. In May 2020, Thurner et. al. [25] first found a constant case rate as a non-special solution of a disease transmission model. These authors have considered a (small-world type) network of people in the shape of a circle with every lattice point being connected to its nearest neighbours. They have then connected some of the lattice points to additional points, located far away on the circle. This accounts for the fact that people tend to interact frequently with their families and much less frequently with outsiders. In other words, Thurner et. al. differ from classical models by accounting for heterogeneity in people's interaction patterns. They find that if the average degree D of the network remains below a critical value then the infection spreads at a constant rate before dying out, while if the average degree exceeds this value then the epidemiological curve resembles the bell-shaped or wave solution of ODE models. They have calculated an approximate expression for the critical degree D_c in terms of the transmission probability and the transmissibility interval.

While Ref. [25] appears to be a significant advance relative to the classical models, there are some overlaps with these models as well. For example, the nearest-neighbour links are interpreted as family ties while the long-distance bridge links are interpreted as "social contacts outside the local community (family)". Thus, the constant-rate solution appears to hold only when there is a hard lockdown. Indeed, in a note added to the Supplement during proof (July 2020, by which time second waves had broken out in a lot of places), the authors mention that "many countries have (at least partially) taken back many NPIs, resulting in a (seemingly exponential) resurgence of daily infections ….. [which] is maybe not yet the 'second wave' but just the logical [fallout of the] reduction of social distancing and [consequent] increase of D, d [transmissibility duration] and ε [probability of a bridging link existing between two distant lattice points]". Indeed, the statement that in the network model, the epidemic is contained below a certain average network degree D_c and evolves naturally above it is substantively the same as the statement that in the DDE model [9], the epidemic is contained below a critical interaction rate (for which $R_0 = 1$) and evolves naturally above it. The nature of the containment solution is different in the two models (constant rate in Ref. [25] vis-a-vis decaying rate in Ref. [9]), but containment and wave solutions are still separated by a change in a population-averaged parameter. It is worth noting that a rebuttal [26] exists for this study; the rebutters allege that the Thurner et. al. model generates constant case rate only in a tiny region of parameter space where R_0 is very nearly unity and hence that the constant solution is no more generic than it is in the classical models.

In February 2021, Tkachenko et. al. [27] conjectured that the plateaus and waves are generated as a result of temporal heterogeneity in interaction rate, rather than the heterogeneity in our interaction patterns. Temporal variation refers to the fact that a person will have a high interaction rate with others on some days, for example when s/he attends a party, and a lower interaction rate on other days, for example when s/he stays at home. To some degree at least, we would expect such fluctuations to get smoothed out when the case counts are high, and indeed their final model equation is quite similar to the conventional S-I-R model. Nielsen et. al. [28] have focussed on heterogeneity in people's infectivity i.e. superspreading incidents. Manrubia and Zanette [29] have contended that plateau solutions ($R = 1$) are primarily the result of individuals' risk-averse behaviour rather than a heterogeneity effect. The plateaus have also been considered in terms of the "critical slowing down" phenomenon, in a model-agnostic manner [30] as well as with reference to an S-I-R model with continuously time-varying parameters [31].

We believe that heterogeneity in interaction patterns remains the most plausible hypothesis behind the non-classical epidemic trajectories. After all, most people socialize within narrow groups of family and friends – interactions outside this set are considerably rare. Therefore, we build a transmission model which starts from this basic premise. Our model is ultimately deterministic although it is based on probabilistic concepts. We find that it is capable of predicting plateaus and waves as well as unexpected transitions from the first state to the second.

---- o ----

MATHEMATICAL MODEL

§4. No vaccination. Before starting the model development, we mention that in this entire Article, we shall ignore vaccination. This is because COVID-19 plateaus and waves occurred in many countries when zero or small fractions of their populations were vaccinated. Multiple sudden surges occurred in America and Europe in mid-late 2020 when there were no vaccines available. India's second wave started when only about 10 percent of the population had received their first dose. Even now, USA has some distance to go in terms of vaccination coverage while Taiwan and Japan are still further back. Therefore we leave the consideration of this intervention for a future study.

§5. Qualitative understanding. It is our observation that in late 2020 and early 2021, in many regions of India and USA, high levels of socioeconomic activity could be sustained for weeks on end while cases remained on a plateau – restaurants, cinema halls, places of worship as well as public transportation remained open without triggering case explosions. Similar phenomena must have occurred in other countries as well. We first argue how this can even be possible. Masking is literally absent in eateries, and we are aware that some violations have also been occurring in cinema halls, public transportation and other places. We hypothesize as follows. Firstly, the vast majority of symptomatic people go into quarantine or at least refrain from interacting in society, if for no other reason than that public sneezing or coughing makes one an immediate target of suspicion. Hence, almost all transmissions occur from asymptomatic or latent (pre-symptomatic) cases via speech and breathing. Masks render these mechanisms almost incapable of transmission [32], so the majority of spreading occurs in unmasked settings. Even there however, breathing and speech carry the infected droplets over a relatively short distance (speech farther than breathing). Thus, if the virus is present in a restaurant, it is very much likelier to spread among people seated at the same table than to jump from one table to the next (or to saturate the restaurant and infect everyone inside). Similarly, mask fault in a cinema hall might also infect only the nearest neighbour and not someone seated far away. In summary, our first hypothesis is that the overwhelming majority of transmissions occurs during close, unmasked interactions.

It is logical that a person will indulge in such interactions only with his/her family members and friends and not with strangers. Thus, even if Alfa goes to 20 restaurants in two months, it is quite possible that her dining companions on all these occasions will be a subset of Bravo, Charlie, Delta and Echo. (We use the codewords from the ICAO phonetic alphabet rather than Alice, Bob etc to denote random peoples' names since (*a*) they are country- or culture-independent, (*b*) the names are suggestive of gender, avoiding the use of cumbersome gender-neutral pronouns, and (*c*) among the first four candidates, two – Bravo and Charlie – are naturally male while two – Alfa and Delta – are naturally female which automatically ensures gender symmetry). Similarly, Bravo's companions might be Alfa and Delta as well as Foxtrot while Charlie's might be all the previous plus Golf. Our second hypothesis comes now. We posit that, starting from a random person, if we keep extending these links of close contacts, then it is extremely unlikely that the chain will continue all the way upto the region's last inhabitant. On the contrary, sooner rather than later, the links will cover no new person and the cycle will close. This will give us a group of social contacts with dense links amongst each other and no (or very few) links outside. We call this group a **cluster**. At a small office for example, all employees might belong to a single cluster; at a large company, there will be multiple clusters with the people in each cluster possibly belonging to the same rank or payscale, or sharing the same office space. By the nature of clusters, when the virus enters a cluster it will spread rapidly among its members, but will find it much more difficult to infect someone outside the cluster.

The two hypotheses – primary spreading from close interactions and these being confined within clusters – can explain why it is possible to have low disease prevalence even with public transport, entertainment venues and places of worship open at full tilt. At recreation venues we interact primarily within our cluster and facilitate intra-cluster transmission. As regards places of worship, we go there either alone or with family, pray and come back. Inside a holy place we do not spend time socializing with all and sundry. Finally, in public transit as well, we by and large keep to ourselves, or interact with members of our cluster if we are going together to an entertainment venue. Thus, case counts remain low all through these activities as the virus remains primarily confined within a few clusters.

Of course, all transmission cannot be occurring inside the clusters; if that were the case then the epidemic would not perpetuate. So we must now examine the mechanisms by which the disease can jump from cluster to cluster. There are two ways this can happen. The first is *unintentional* – during necessary activities like shopping, local travelling or working, an unknowing case's mask might happen to slip off just when a potential target is close by, or we may touch a contaminated surface/object and then our face, or the virus might just jump across a pair of masks etc. The second mechanism of inter-cluster jumping is social occasions where we *deliberately* interact outside of our cluster. For example, the invitees at a marriage gathering might include multiple families who hate each other and normally do not meet (and even less so in pandemic times). Similarly, a birthday bash hosted by a well-to-do IT sector company employee might feature half the staff in attendance, an occurrence which would not take place at a casual entertainment venue. At events like this, adherence to COVID-appropriate behaviour tends to be low; worse still, social norms require us to actively interact with many people present at the gathering and not just with our family members or close friends. Thus, a case present at such a gathering will likely transmit the disease to several others not belonging to his/her cluster.

For the purposes of model-building we need to distinguish between household transmission and transmission among friends. In the former situation, due to constant contact, a case will spread the disease to all household members as soon as s/he turns infectious. In the latter situation however, it will take a finite time to fully infect a group – for example, the at large case Alfa might dine with Bravo and Charlie on one day and infect them both, Charlie will fall sick next week and go to a movie with Echo and so on; it is unlikely that everyone from Alfa to Xray will be simultaneously present at an eatery or movie theatre and get infected in one fell swoop. To achieve this distinction, we must make two assumptions. Firstly, we decouple households and clusters, taking the latter to include only those close contacts with whom a person does not live together. Secondly, we assume that, per household, there is exactly one member who is socially active i.e. part of a cluster, while the other members are completely cautious. For example, in a family where a young working woman lives with her retired parents, the latter two might not step out of the house during the pandemic while the woman goes to work and does the shopping etc. A cautious family member can also be a person who is not socially inactive but rigorously adheres to COVID-appropriate behaviour all the time. In either case, the only way that the cautious person catches corona is if the active person catches it first; the active one can catch it through intra- or inter-cluster transmission.

In view of the above, we build our mathematical model taking into account a four-tiered viral transmission process, as follows.

- **Household transmission :** As soon as the socially active member of a household contracts the virus, the cautious ones follow suit.
- **Cluster transmission :** When one member of a cluster gets infected, the others also gradually fall sick. The rate at which this happens and the fraction of people in the cluster who contract the infection are determined by the properties of the virus and the interaction rate of the cluster members among each other.
- **Unintentional cluster transition (UCT) :** Note that this is cluster *transition* and not *transmission* – a jump from one cluster to another. These are events like the accidental mask slippage in a shop or a bus mentioned above. In this category we also include the events where a person makes a new acquaintance outside his/her cluster and starts socializing with him/her – we expect that events of this kind will be rare overall.
- **Socializing external to cluster (SEC) :** These are events like the wedding or party mentioned above. Organized gatherings where people from different clusters interact with each other also belong to this category.

Intuition says that SEC events might be very dangerous from a public health perspective, since they are almost designed to facilitate large-scale inter-cluster transmission. We shall now verify this intuition through the mathematical model.

§6. Quantitative model development. Our model treats time to be discrete and measured in days i.e. we construct a map rather than a flow. The population is also discrete as in an agent-based model, rather than continuous as in an ODE/DDE model. However, our model is ultimately deterministic rather than stochastic. The treatment in this Section is somewhat involved, so we have presented a summary recap in §7. Let us consider a city (see later for more clarification) of total population N, with everyone initially susceptible. Let the N people belong to N_1 households where $h = N/N_1$ is the average household size. With the assumptions of §5, there are N_1 people who belong to social clusters and can contract the virus directly; once this happens, each of them transmits to the $h - 1$ other members of their household (or family). Thus we can focus on the disease dynamics only among the N_1 socially active people, and add on the family cases at the end. The values we have chosen are $N = 302400$ and $h = 3$ so that $N_1 = 100800$. Our N is chosen to enable a comparison with the results in Ref. [9] where we have used Notional Cities of a similar population, and also to ensure that individual runs get completed in a reasonable time-frame on a commonly available laptop computer; the average household size of three is reasonable.

In the next step, we divide the N_1 socially active people into N_C clusters. Here, we assume that all clusters have the same size s, so that $N_1 = sN_C$. The choice of 100800 for N_1 rather than exactly 10^5 ensures that N_1 is divisible by a lot of two-digit numbers; this enables easy variation of s. The value we have used here is $s = 24$, so that $N_C = 4200$. To fix the intra-cluster dynamics, we now need some elementary characteristics of viral transmission. We assume that the serial interval is 5 days [33,34] i.e. 5 days elapse between a primary case and a secondary case's turning transmissible. We also assume that, once a person turns transmissible, s/he spends 3 days at large before recovering. The 3 day transmission interval is an average of the weeklong asymptomatic period and the approximately 1 day latent infectious period before a symptomatic case seeks quarantine [35]. The assumption of recovery (rather than isolation, hospitalization or death) after three days is simply for analytical tractability.

Our next necessary piece of information is that the basic reproduction number R_0 of COVID-19 is somewhere between 2 and 5, depending on the viral strain etc [36,37]. This means that in the absence of interventions, one person spreads the disease to between 2 and 5 people. For intra-cluster transmission we assume a value of $R_0 = 2.5$ (later we also consider $R_0 = 4$), and construct an extremely heuristic pathway for progress of the infection through the cluster. The R_0 and serial interval mean that 5 days after the first person in a cluster is exposed, s/he infects 2.5 others – since fractional infections in a cluster of 24 make no sense, we round it off to 3. Then, 5 days later, each of these 3 cases exposes 2.5 people for a total of 7.5

exposures; since by this time 4 people are already infected/immune, the number of actual cases at this step should be 7.5×(20/24) which is 6.25, rounded down to 6. Thus, 5 days after the 3 cases, the cluster develops 6 more cases. Similarly, at the next step, these 6 cases expose 15 further people and the susceptible fraction 14/24 among them turn into cases. This gives us 8 fresh cases. Analogously, these 8 generate 5 more cases after 5 days and then these 5 infect the last one after 5 further days. From this argument we get the **cluster sequence** [1; 3; 6; 8; 5; 1]. The i^{th} element of this sequence indicates the number of people who contract the infection i serial intervals after the first person in the cluster is exposed. It is probably unrealistic that every single person in a cluster will get infected; in reality, we expect a lucky person or two to be spared. Taking this into account, we modify the sequence to [1; 3; 6; 7; 5; 1], which leaves one person uninfected at the end of the cluster-level outbreak. The entire concept of cluster sequence is heuristic, and is necessary only to get a deterministic model – we discuss the implications of this sequence in §16.

The existence of a cluster sequence means that we can define each cluster to be in one of two states : susceptible if all members of the cluster are susceptible, and insusceptible as soon as the first member has been exposed, and for ever after (we assume permanent immunity, which seems to be valid so far [38,39]). During the 30 days it takes for the 23 cluster members to be infected, we treat the cluster as a whole to be insusceptible – we assume that any further exposure of cluster members during this period does not change the intra-cluster infection pattern.

At this point we can introduce the variables in the model. We count a person as a case on the day that s/he first turns transmissible. Let y_i be the cumulative number of cases occurred upto and excluding day #i and let Δy_i be the additional cases occurring on day #i itself, so that $y_{i+1} = y_i + \Delta y_i$. Similarly, let z_i be the cumulative count and Δz_i the daily count of clusters which are seeded upto and on day #i respectively. y and z are obviously not independent; by definition of the cluster sequence if $\Delta z_{100} = 1$ then Δy_{105} gets a contribution of 1, Δy_{110} gets a contribution of 3, Δy_{115} gets a contribution of 6 and so on. We say "gets a contribution of" rather than "equals" because Δy_{105} will also carry contributions from clusters which have been seeded on days #75, #80, #85, #90 and #95. In addition we have the near-dummy variables for family cases f_i and Δf_i; since every active member will infect his/her two household members as soon as s/he turns infectious, and since the incubation period is 5 days, we have $\Delta f_{i+5} = 2\Delta y_i$ and $f_{i+5} = 2y_i$.

Having accounted for the first two transmission mechanisms of §5, we now start work on the latter two i.e. the UCT and SEC modes. Our ultimate question is : given the case histories upto day #i and the interaction patterns for UCT and SEC, what is the number of susceptible clusters which get seeded on day #i ? This number will enable us to move one step forward in time i.e. from day #i to day #i+1. Since the phenomena involved are probabilistic but the model is deterministic, we must calculate the expectation value. We now demonstrate how to do this. Our baseline premises are : (*a*) on any given day all the N_1 active people are equally likely to participate in UCT and SEC events, and (*b*) at both these events, clusters are mixed at random i.e. a case from cluster #101 is as likely to infect a member of cluster #102 as a member of cluster #2302. Assumption (*b*) is the equivalent of homogeneous mixing in classical epidemiological models but at the cluster rather than the individual level. These assumptions ensure that the most realistic domain of validity of our model is a city but not a larger region, such as a state or country or the world. A marriage function or birthday bash held in one city is likely to have maximum guests from the same city, and few if any guests from another city in the same state, another state in the same country, or another country. Since all cases remain at large for three days, the total number of at large cases present on day #i is $\alpha = \Delta y_{i-1} + \Delta y_{i-2} + \Delta y_{i-3}$. SEC transmission is conceptually more concrete than UCT so we take that first. Let n_S (constant in time) be the total number of people who participate in SEC events every day – it does not matter whether it is a single gathering of size n_S people or fifty separate gatherings adding up to total n_S people which are taking place. We assume that if everyone is susceptible then one case attending an SEC event spreads the disease to m_S people at the event. We use the value $m_S = 2$.

The number of cases attending SEC events on day #i can be anything between 0 and α. The probability that the number is exactly k is

$$P(k) = \frac{{}^{\alpha}\mathrm{C}_k \; {}^{N_1-\alpha}\mathrm{C}_{n_S-k}}{{}^{N_1}\mathrm{C}_{n_S}} \quad , \tag{1}$$

where the combination function ${}^{n}\mathrm{C}_r$ denotes the number of ways of selecting a team of r players from a pool of size n. Given that there are k cases participating in SEC events, by the model assumptions they transmit an infective dose of virus to $\beta = 2k$ people (we say "transmit an infective dose" rather than "transmit the disease" since the recipient of the viral dose contracts the infection if and only if s/he is susceptible). The next question is, how many distinct clusters do these β people belong to ? This is relevant because if five recipients of infective viral dose belong to five different susceptible clusters then there will be five cluster sequences manifest over the next 30 days while if all the recipients happen to belong to the same susceptible cluster then (by the model assumptions) there will be only one cluster sequence manifest during this period. The possible number of clusters can range from between $\lceil \beta / 24 \rceil$ (smallest integer greater than or equal to $\beta/24$) and β, and we must ask what is the probability that the β people belong to exactly b clusters. In the Appendix we show that the answer is

$$P\left(b \mid \beta\right) = \frac{{}^{N_C}\mathrm{C}_b \; {}^{\beta-1}\mathrm{C}_{b-1}}{{}^{N_C-1+\beta}\mathrm{C}_{\beta}} \quad . \tag{2}$$

There is one more conditional probability to be taken care of. We know that b clusters are infected at SEC events, but how many of them are actually susceptible ? As the disease progresses, a higher and higher number of the b clusters will actually be insusceptible ones. On day #i there are z_i insusceptible clusters and $N_C - z_i$ susceptible ones. The probability that among b randomly selected clusters, exactly j are susceptible is

$$P\left(j \mid b\right) = \frac{{}^{N_C-z_i}\mathrm{C}_j \; {}^{z_i}\mathrm{C}_{b-j}}{{}^{N_C}\mathrm{C}_b} \quad , \tag{3}$$

and all the conditional probabilities are on the table.

Now, the quantity of interest is the expectation value of the number of new susceptible clusters seeded on day #i. This can be calculated as

$$E_S\left(\Delta z_i\right) = \sum_j jP(j) \quad , \tag{4}$$

where the subscript S reminds us that this is the expected number of clusters seeded during SEC events and $P(j)$ is the total probability that j susceptible clusters are seeded on day #i. So far we have the partial probabilities, • conditional $P(j \mid b)$: given that an infective dose of virus is introduced into b clusters, the probability that j of them are susceptible, • conditional $P(b \mid \beta)$ [or equivalently $P(b \mid k)$ since $\beta = 2k$] : given that $2k$ people receive infective viral doses, the probability that they belong to exactly b clusters, and • absolute $P(k)$: the probability that k cases are actually participating in SEC events on the day in question. Thus, given a pair k,b the probability of j susceptible clusters' being seeded is $P(j \mid k,b) = P(k)\,P(b \mid k)\,P(j \mid b)$. The total probability $P(j)$ will be this summed over all possible k and b. k runs from 1 to a maximum of α while b runs from 1 to a maximum of β. Thus we have

$$\begin{aligned} P(j) &= \sum_{k=1}^{\alpha} P(k) \left(\sum_{b=1}^{\beta} P(b \mid k) P(j \mid b) \right) \\ &= \sum_{k=1}^{\alpha} \frac{{}^{\alpha}\mathrm{C}_k \; {}^{N_1-\alpha}\mathrm{C}_{n_S-k}}{{}^{N_1}\mathrm{C}_{n_S}} \left(\sum_{b=1}^{\beta} \frac{{}^{N_C}\mathrm{C}_b \; {}^{\beta-1}\mathrm{C}_{b-1}}{{}^{N_C-1+\beta}\mathrm{C}_{\beta}} \frac{{}^{N_C-z_i}\mathrm{C}_j \; {}^{z_i}\mathrm{C}_{b-j}}{{}^{N_C}\mathrm{C}_b} \right) \end{aligned} \quad . \tag{5}$$

Now, we must implement the sum (4). For each b, j can run from 1 to b and we can pull the summation over j inside the second of the two sums in the above right hand side. Doing so gives us the expectation value

$$E_S\left(\Delta z_i\right)=\sum_{k=1}^{\alpha}\frac{{}^{\alpha}\mathrm{C}_k\ {}^{N_1-\alpha}\mathrm{C}_{n_S-k}}{{}^{N_1}\mathrm{C}_{n_S}}\left[\sum_{b=1}^{\beta}\frac{{}^{N_C}\mathrm{C}_b\ {}^{\beta-1}\mathrm{C}_{b-1}}{{}^{N_C-1+\beta}\mathrm{C}_{\beta}}\left(\sum_{j=1}^{b}j\frac{{}^{N_C-z_i}\mathrm{C}_j\ {}^{z_i}\mathrm{C}_{b-j}}{{}^{N_C}\mathrm{C}_b}\right)\right] \quad , \tag{6}$$

and the contribution of SEC events has been determined.

In a similar manner, we can calculate the contribution of UCT events to inter-cluster spread. Let n_U be the number of people participating in UCT events every day – these are the people who visit crowded markets, travel on crowded buses and trains etc. Let a case present at an UCT event transmit an infective dose of virus to m_U targets. Unlike for SEC, we expect that on the average m_U will be less than unity, in which case we can also interpret m_U as the probability P_U that a case successfully transmits an infective dose to a target at an UCT event. Since this probability is a parameter present in classical epidemic models as well, we use P_U rather than m_U. However, for the bulk of the calculation, we treat it just like m_S, and keep open the possibility that the parameter value might exceed unity (in which case m_U is the only interpretation which makes sense). Then, we can repeat the argument for the SEC contribution. Given that there are k cases at large on day #i, the expected number of people who receive an infective dose is kP_U. This is the equivalent of β, with one difference; while $\beta = 2k$ was an integer by definition, kP_U will in general not be one, and the entire calculation is cast in terms of integers. To tackle this, we introduce the **roundoff function**. In this function, we define γ to be the integer nearest to kP_U; since this can cause accumulating errors with increasing i, we also retain the difference between γ and kP_U and keep incrementing it with every successive i. When this difference exceeds +1 or –1, we include the correction to γ. This ensures that rounding errors do not accumulate in time. Roundoff apart, everything else is the same. Given that γ people have received an infective dose, we can calculate the probability that they belong to b different clusters and then that j of them are susceptible, and arrive at

$$E_U\left(\Delta z_i\right)=\sum_{k=1}^{\alpha}\frac{{}^{\alpha}\mathrm{C}_k\ {}^{N_1-\alpha}\mathrm{C}_{n_U-k}}{{}^{N_1}\mathrm{C}_{n_U}}\left[\sum_{b=1}^{\gamma}\frac{{}^{N_C}\mathrm{C}_b\ {}^{\gamma-1}\mathrm{C}_{b-1}}{{}^{N_C-1+\gamma}\mathrm{C}_{\gamma}}\left(\sum_{j=1}^{b}j\frac{{}^{N_C-z_i}\mathrm{C}_j\ {}^{z_i}\mathrm{C}_{b-j}}{{}^{N_C}\mathrm{C}_b}\right)\right] \quad . \tag{7}$$

The expected total number of susceptible clusters seeded through SEC and UCT events is

$$E\left(\Delta z_i\right)=E_S\left(\Delta z_i\right)+E_U\left(\Delta z_i\right) \quad . \tag{8}$$

Since this will in general be a fraction, we again use the rounding off procedure described above, approximating it to the nearest integer and compensating for the error. Thus a roundoff on (8) gives Δz_i. At once we can implement Δz_i times the cluster sequence over the next 30 days. Finally, we do $y_{i+1} = y_i + \Delta y_i$ and $z_{i+1} = z_i + \Delta z_i$ to complete one iteration of the map and move from day #i to day #i+1.

§7. Summary of the model. The last Section describes the procedure of the model in its full elaboration. We now present it in a more concise form, especially for those who may have skipped the details. Since the key feature of our model is the incorporation of social clusters, we call it the **cluster seeding and transmission** or **CST model**. First, in Table 1 we present the variables and in Fig. 3 the parameters involved, as well as the values of the latter which correspond to a default or baseline solution.

Variable	**Significance**
i	Discretized time, measured in days
y_i	Cumulative number of socially active cases upto and excluding day #i
Δy_i	Number of new socially active cases cropping up on day #i
z_i	Cumulative number of insusceptible clusters upto and excluding day #i
Δz_i	Number of new clusters turning insusceptible on day #i
f_i	Cumulative number of cautious household cases upto and excluding day #i
Δf_i	Number of new cautious household cases cropping up on day #i

Table 1 : *Variables in the CST model.*

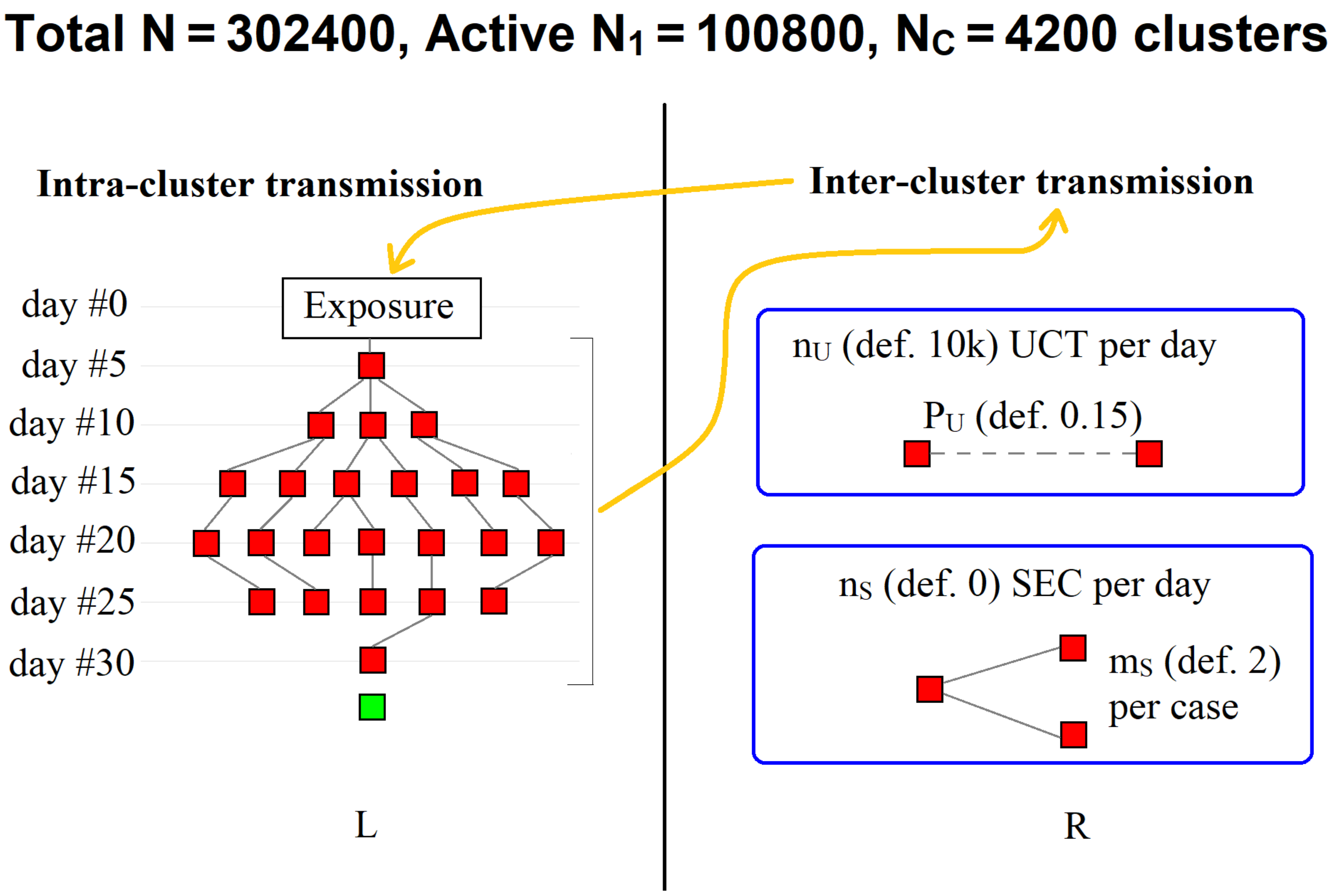
Total N = 302400, Active N₁ = 100800, N_C = 4200 clusters
Intra-cluster transmission
Inter-cluster transmission
Exposure
day #0
day #5
day #10
day #15
day #20
day #25
day #30
nU (def. 10k) UCT per day
PU (def. 0.15)
nS (def. 0) SEC per day
mS (def. 2) per case
L
R

Figure 3 : *Schematic representation of the parameters in the CST model. The left panel depicts the cluster sequence (in a very heuristic manner) while the right panel shows UCT and SEC events (red squares denote cases). The yellow arrows indicate how the two transmission modes are interlinked. "Def." indicates default value.*

The algorithm itself is as follows.

Subroutine roundoff

- This operates continuously during the main loop on i. To start, define an additional variable total_err whose starting value is zero.
- For every i, after rounding the target variable to the nearest integer, calculate the error and add it to total_err. If this latter exceeds +1 respectively −1, then increment respectively decrement the target variable value by 1 and decrement respectively increment total_err by 1.

Main routine

- Initialize all variables with zero values. This is important since there are additive steps. Implement the initial conditions.
- Perform a big loop over i with the following steps
    - Define $\alpha = \Delta y_{i-1}+\Delta y_{i-2}+\Delta y_{i-3}$; break loop if α equals zero beyond the seeding phase
    - Define δ = roundoff(km_S) for each k and calculate E_S using (6). Note that rounding off is not necessary if m_S is taken as the default value 2, but is required for a more general, non-integer value
    - Define γ=roundoff(kP_U) for each k and calculate E_U using (7)
    - Use (8) to calculate E and define Δz_i = roundoff(E)
    - Set $\Delta y_{i+5} = \Delta y_{i+5}+1$, $\Delta y_{i+10} = \Delta y_{i+10}+3$, $\Delta y_{i+15} = \Delta y_{i+15}+6$ etc following the cluster sequence
    - Set $y_{i+1} = y_i+\Delta y_i$, $z_{i+1} = z_i+\Delta z_i$ to complete one iteration of the loop
- Add on the household cases arising from socially active cases
- Prepare plots as necessary

The default initial condition is that eight clusters are seeded on the first day. The Matlab code implementing the algorithm outlined above is available on demand, as described in the Data Availability Statement at the end of the Article. For computational convenience, we introduce a parameter $k_{\max}$ in the sum over k in (6) and (7). Ideally, this sum should run all the way upto α. However, with n_U and n_S both being significantly smaller than N_1 (typically 1-10 percent), the probability that α or nearly α cases will all be participating in UCT or SEC events on the same day will be minuscule. Hence, we can define a cutoff $k_{\max}$ above which we shall just treat this probability to be zero, and ignore the error. The computational time and effort increase very rapidly with increasing $k_{\max}$ so it is worthwhile to choose its value judiciously. For all display results here, we use $k_{\max} = 80$.

Finally, the expressions (6) and (7) must be inputted to the computer in a special manner; we describe this in the Appendix.

---- o ----

SIMULATION RESULTS AND ANALYSIS

§8. The default solution. This corresponds to the solution of the model with the default parameter values from Fig. 3. We call these values defaults not because they have been obtained from any data fits (see §16 for a discussion of this issue) but because variation of each parameter on either side of the default leads to different kinds of interesting behaviour. The time trace of the epidemic with these values is shown below. In all these plots, we shall show the cumulative case count as a black line associated with the right hand y-axis and the daily case rate as grey bars associated with the left hand y-axis.

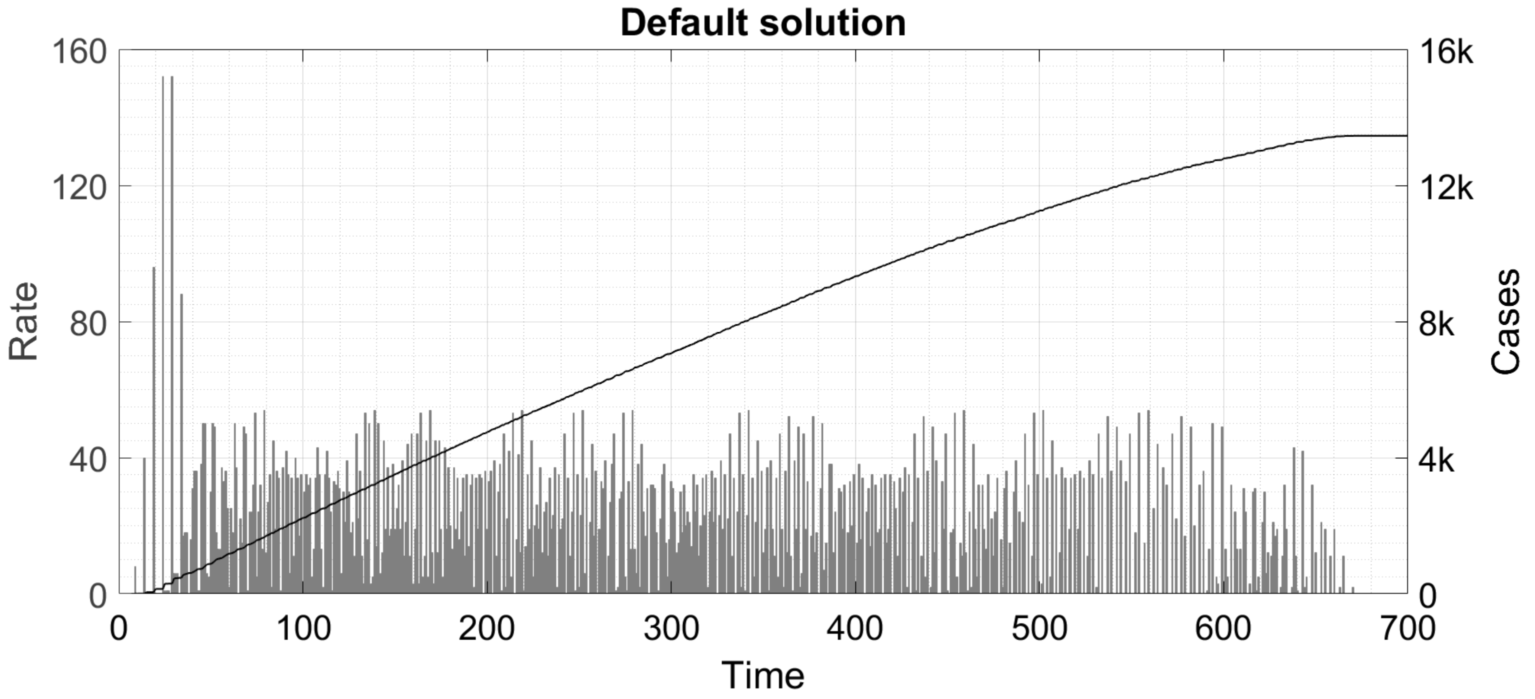

Figure 4 : *The default solution showing constant daily case rate. The symbol ‘k’ denotes thousand.*

We can see that the epidemic continues for a long time at almost constant daily case rate before eventually dying out. This is the plateau solution.

§9. Variation of P_U. Here we present the case trajectories as P_U (the probability that a UCT event actually leads to a transmission) is varied. Keeping all other parameters fixed at their default values, we consider three representative values of P_U and display the time trace of the epidemic in the three panels of Fig. 5 below. We have used the same x- and y-axis scalings in all panels so that the contrasts may be visually apparent.

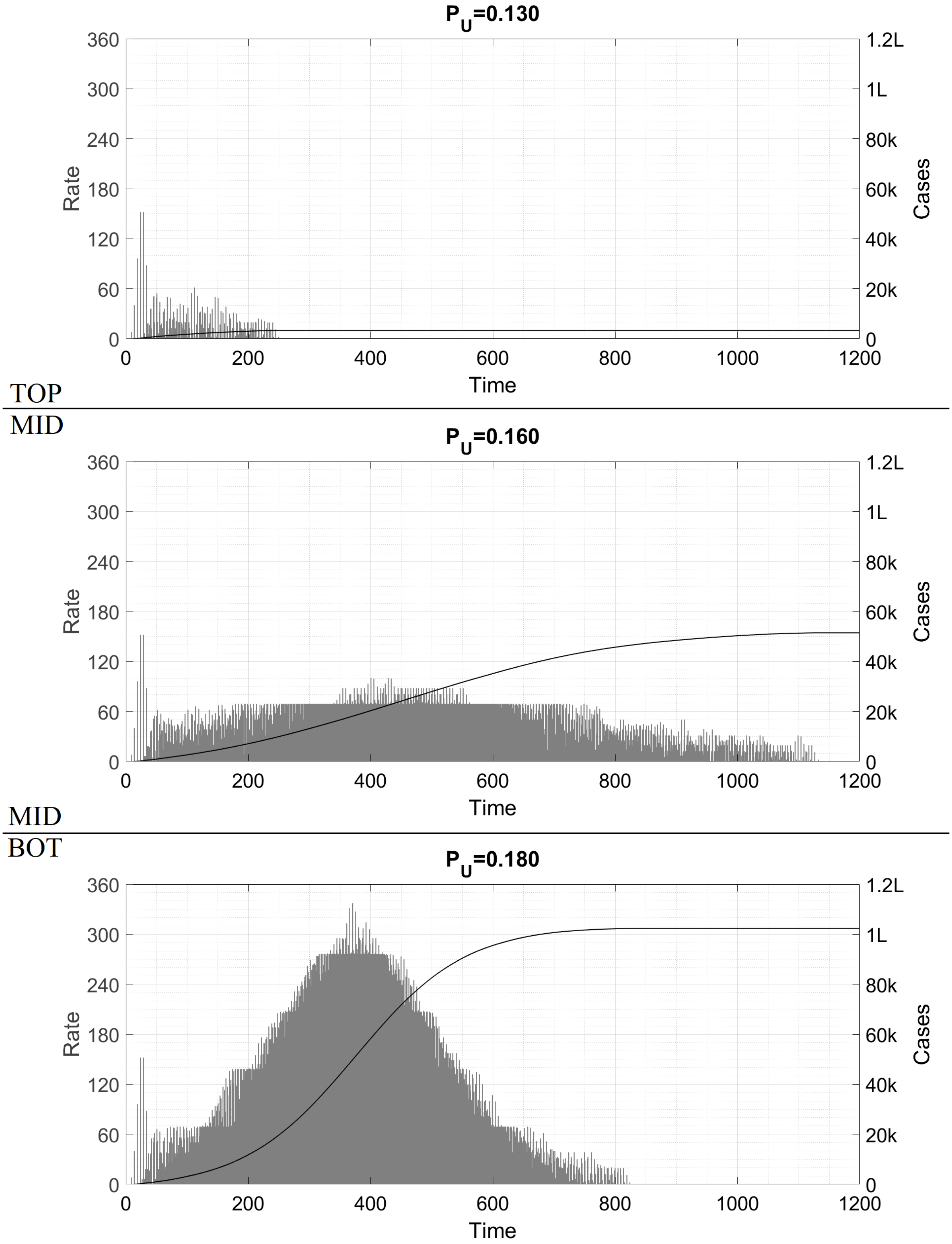

Figure 5 : *Time traces of the epidemic as* P_U *is varied. We can see increasing caseload and peak rate with increasing* P_U*. The symbol 'k' denotes thousand and 'L' hundred thousand.*

Between the three panels, we can see elimination, plateau as well as wave solutions. The behaviour obtained by increasing n_U while keeping P_U constant is remarkably similar, and we do not repeat the figures here.

§10. Classical instability. The plateau, Fig. 4, has first been reported by Thurner et. al. [25]. We note that the chosen default value of 10000 for n_U is realistic since our population is assumed to consist of 100800 socially active people. On average, each person participates in a UCT once every 10 days – this factors in that UCT requires mask slippage, separation minima infringement and similar lapses which do not happen to every person during every necessary interaction. The probability P_U of Fig. 5 is a parameter which is closely related to classical epidemiological models – in the S-I-R model, it is accommodated into the force of infection while in the DDE model [9] it enters as a multiplicative factor in the per-case spreading rate (this is defined as $m_0 = q_0P_0$ where q_0 is the interaction rate and P_0 the transmission probability). It is expected that a lower P_U will lead to a better outcome and vice versa, and Fig. 5 shows that this is indeed the case. Upto this point, our model agrees fully with classical models.

The first point of difference also comes up in Fig. 5. For, while at $P_U = 0.130$ the epidemic dies down in time after an initial phase of constant rate (probably caused by the strong seeding), at $P_U = 0.160$ the epidemic plateaus, just as it did with $P_U = 0.150$. Thus, the constant solution is actually valid for a range of P_U instead of just one value of P_U as happens in a classical model (the special value corresponding to $R_0 = 1$). At sufficiently high P_U however, the constant solution is no longer seen, and is replaced by a wave solution. By iterating the code, we have found that the plateau remains valid in the approximate range P_U belongs to [0.135, 0.165]. A more rigorous determination of the stability boundaries – based on a formal mathematical criterion rather than eye estimation – is left by us for a future study. We shall say that $P_U < 0.135$ corresponds to a **stable** region of parameter space, $0.135 < P_U < 0.165$ to a **neutral** region of parameter space, and $P_U > 0.165$ to an **unstable** region of parameter space. These words are derived from the classical theory of epidemic dynamics, where the corresponding regions of parameter space result in the disease-free equilibrium (for the ODE models) or an arbitrary equilibrium (for the DDE model [9]) being stable, neutral and unstable respectively (the neutral region in these models actually has measure zero). Within the neutral range, the value of the constant rate and the duration of the epidemic both increase with increasing P_U. In the unstable region, an increase in P_U causes the wave height to increase and the duration to decrease, in such a manner as to increase the cumulative caseload. This conclusion again agrees with the DDE model [9] and other classical models.

Note also that stable, neutral and unstable regions of parameter space are all defined with respect to the prior infection level; for example, a parameter set which is unstable for fully susceptible population might be neutral for 25 percent initial infection level and stable for 50 percent infection level. Thus, in Fig. 5-mid, the mode of operation changes from neutral to stable at approximately 800 days while in Fig. 5-bot, the operation changes from unstable to neutral/ stable at approximately 400 days (in the classical model, we would say that R decreases across unity at this point). Practically, the existence of a constant solution over the entire neutral range of parameter values complicates the epidemic management process as it implies that, seeing a constant solution in reality, we are not aware of how close we are to an instability. However, even after passing the instability, the system behaviour with increasing P_U remains tractable – a small increase in P_U causes only a gradual increase in case rate, which gives the authorities enough time to recognize the instability and reintroduce a higher level of NPI. This feature is again shared with the DDE model, where a small increase in R across the critical value of 1 causes a small increase in the case rate. Thus, Figs. 4-5 add no substantively new information beyond what can already be obtained from Refs. [9] and [25].

A comment about the solution in Fig. 5-mid is in order. We can see that this solution is extremely smooth, with identical case rate persisting from one day to the next over very large periods. This is the effect of a fortuitious parameter choice as well as the subroutine roundoff. The choice of P_U happens to be such that, if there is exactly one cluster seeded every day for more than 30 days, then every day thereafter E_U happens to be very close to 1, only a percent or so above it (recall that since $n_S = 0$, $E_S = 0$ also). Thereafter, the roundoff plays its role. For if E_U is 1.01, then the algorithm will count it as exactly 1 seeded cluster, and seed an extra cluster only after 100 days. These extras manifest as the spikes at around day #400. A similar phenomenon might be responsible for the flat top in Fig. 5-bot. In reality, we would not expect to see such smooth plateaus but much more pronounced fluctuations about a constant solution. In the wave solutions of Fig. 6 (next Section), the absolute numbers of clusters being seeded at the peak are much larger, and the reduction of susceptible population is sufficiently fast as to prevent the algorithm from settling into a steady state with constant E every day. Hence, in that situation, we do not see flat peaks but rounded ones.

§11. Variation of n_S. In this Section, the quantity varied is the number n_S of people participating in SEC events every day. Just as in §9, in Fig. 6 we present three plots of case trajectories for different values of n_S while all other parameters remain at their default values. Again, we use constant scalings on the axes across panels, to facilitate visual comparison. The scalings are different from §9 though.

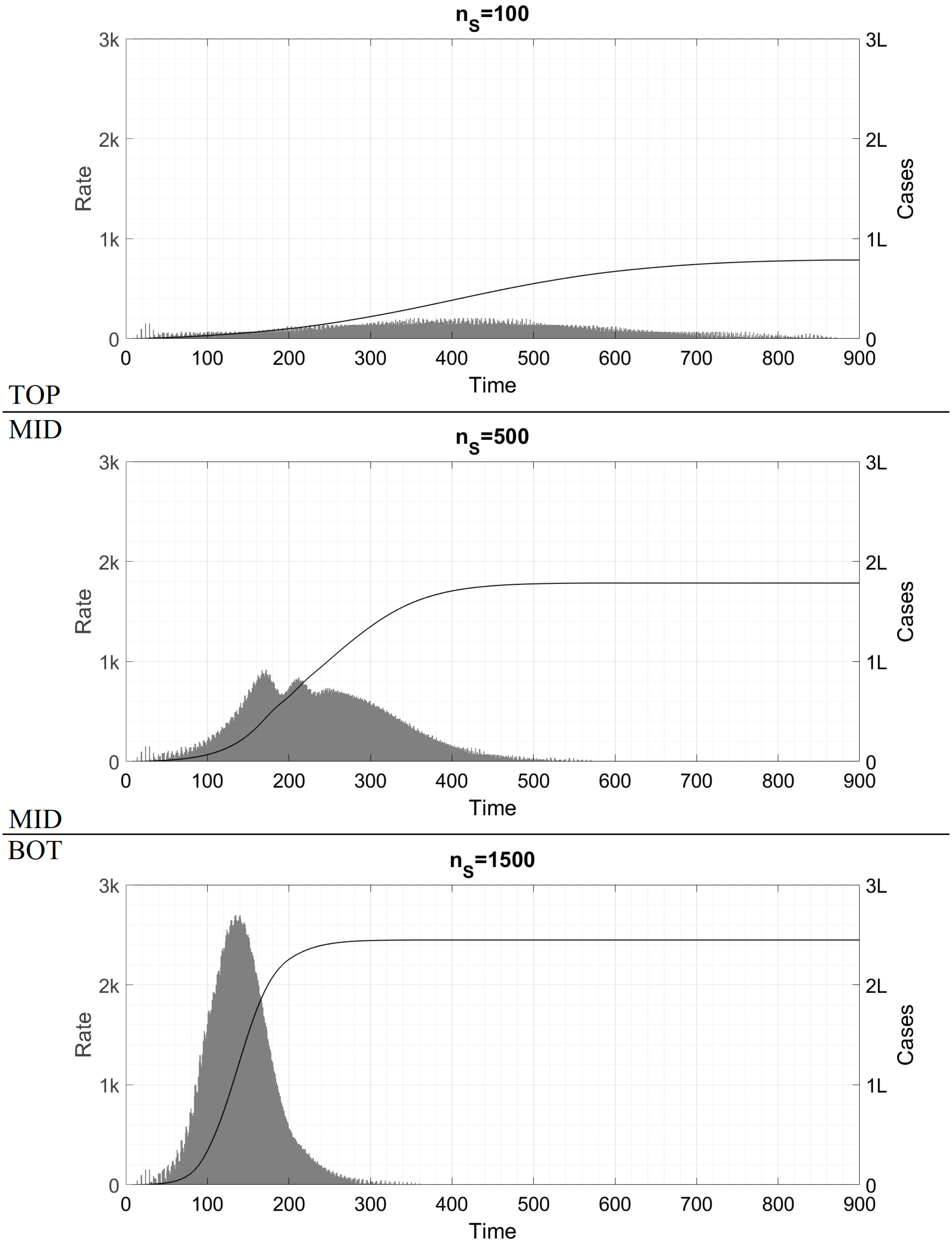

Figure 6 : *Time traces of the epidemic as* n_S *is varied. We can see rapidly increasing caseload and peak rate with increase in* n_S*. The symbol 'k' denotes thousand and 'L' hundred thousand.*

In the last of the three scenarios, the epidemic progresses to what is conventionally known as herd immunity. The behaviour obtained by increasing m_S while keeping n_S constant is qualitatively the same as what we see here, so we dispense with a separate Figure.

§12. Cryptogenic instability. Unlike the instability of Fig. 5 (§10), the one of Fig. 6 has no classical counterpart. We can see that even a small n_S of 100 can destabilize the plateau into a very broad and shallow wave – the cumulative caseload in Fig. 6-top is about six times higher than in Fig. 4. As n_S is further increased, the height and speed of the wave increase dramatically. What makes this instability even more surprising is that from the perspective of classical epidemiological models there is hardly any change at all in the interaction rate between the default solution and the three panels of Fig. 6. In these models, there is only a population-averaged spreading rate which is proportional to the average interaction or contact rate [40]. Let us calculate the average spreading rate for the situations here, when everyone is susceptible. In the default solution with no SEC events, each socially active case spreads the disease to 2 householders and approximately 2.5 cluster members (recall that the cluster sequence of [1; 3; 6; 7; 5; 1] is based on a very rough intra-cluster R_0 of 2.5). In addition, this case participates in UCT on average once every 30 days (since n_U/N is approximately 30) and spreads to an average of 0.15 person there. Thus, during three days (the transmissibility period as per our model), the case further spreads the disease to $0.15\times(3/30) = 0.015$ persons via UCT. Adding these two contributions, we can say that every socially active case spreads the disease to 4.515 persons or, more realistically, 4.5 persons. Since 1 in 3 cases are socially active and the others don't spread at all, on average one case spreads the disease to 1.5 persons. (This is already a surprise since a classical $R_0 = 1.5$ makes us expect a full-blown epidemic and not a constant crawl. But there is more to follow.) Dividing by the transmissibility duration (3 days) gives us an average spreading rate of 0.5 person per day.

Now, consider the situation when n_S people participate in SEC events. By definition, each case present at these gatherings spreads the disease to 2 people, so the population-averaged spreading rate for SEC events is $2(n_S/N)$ per day. For the three panels of Fig. 6, this works out to 0.00067, 0.0033 and 0.01 respectively. These increments are negligible relative to the 0.5 person per day contribution of household and intra-cluster (and UCT) transmission events. Classically, since the contact rate is proportional to the spreading rate, the SEC events add a negligible contribution to the contact rate. There is no way in which a 2 percent increase in contact rate can result in Fig. 4 being transformed into Fig. 6-bot. Hence, this instability does not exist in classical models and we call it **cryptogenic instability**.

Cryptogenic instability is dangerous from the viewpoint of epidemic management for two reasons. Firstly, unless we are aware of its existence, there is no reason to suspect that trips to restaurants and cinema halls constitute COVID-appropriate behaviour while attendance at wedding parties and birthday bashes does not. Secondly, unlike the classical instability where a small breach causes a small case growth, even a small breach here can cause a huge growth. By the time the authorities become aware of the danger and reimpose restrictions, the wave will already have overwhelmed healthcare systems. We also note that in each panel of Fig. 6, the decrease in case rate is achieved by conversion of sufficiently high number of people to the immune state – this is not what happens in practice where lockdowns are imposed to control the spiralling epidemic, and released after the case counts have gone down. Such imposition and relaxation of lockdown can lead to the wave-plateau-wave-plateau type of solutions which we have been seeing in many countries. This however is necessarily modelled by time-dependent parameters and falls outside the ambit of this Figure.

The stark difference from classical formulations arises because in our model, the vast majority of a person's interactions are confined within a small group of people (household and cluster) while the classical models assume that any person's interactions are randomly and uniformly distributed among the entire population (this assumption is called homogeneous mixing). Thus, in our model, clusters start off explosively at $R_0 = 2.5$, but they quickly turn 'herd-immune' and confine the bulk of the infection within themselves. In classical models, an outbreak with $R_0 = 2.5$ can subside only after the entire population is herd-immune i.e. when nearly the entire region has been infected. As a consequence of the cryptogenic instability, a wave can emerge from a quiescent background with only a minuscule change in parameters, while in S-I-R model for instance, it needs significant changes in parameter values to achieve this (discussed, for example, in Ref. [31]).

§13. Effect of initial conditions (IC). A further instance of surprising behaviour is provided by the dependence of the solution on initial conditions. The discrete-population nature of the model means that this dependence is non-trivial. In particular, a sufficiently large seeding caseload is required for the epidemic trajectory to be manifest. For example, if the default solution of Fig. 4 is seeded with four clusters instead of eight, then the epidemic terminates almost immediately instead of continuing at constant case rate. A more dramatic example is shown in Fig. 7. Here, we choose $n_S = 1000$ and stick to the other default parameter values. Extrapolating from Fig. 6, this corresponds to a highly dangerous mode of operation and is expected to cause a huge wave. Instead of starting by seeding eight clusters however, this time we introduce some numbers of external cases Δy_i on days 7, 10 and 13. These external cases are not part of any cluster but they participate in UCT and SEC events and spread the infection to the local cluster members. The panels of the Figure are titled by the vector $[\Delta y_7; \Delta y_{10}; \Delta y_{13}]$. Note that the axis scalings in the three panels this time are not the same.

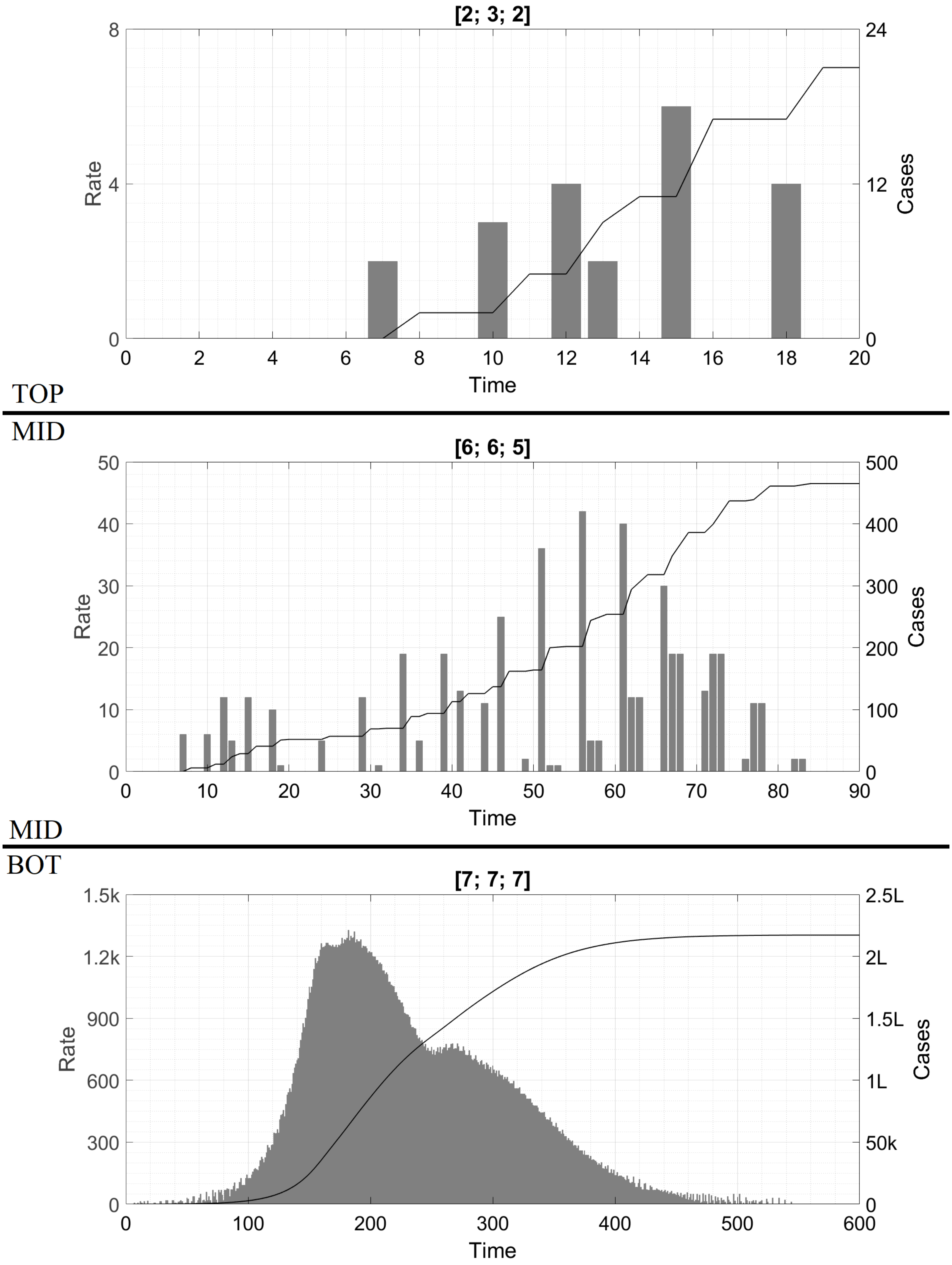

Figure 7 : *Time traces of the epidemic as parameters are held constant but initial conditions are varied. The trajectory depends significantly on the IC used. The symbol 'k' denotes thousand and 'L' hundred thousand.*

We can see that for the smallest IC, not even one cluster is seeded and the epidemic stops at the imported cases. For the intermediate IC, six local clusters are seeded over 50 days before the epidemic runs out of steam. For the largest IC, the epidemic proceeds as we would expect it to. At all seeding vectors totalling 20 cases or more, we found the wave solution; below this threshold, it appeared that the boundaries were blurry. For example, the IC [4; 4; 4] actually led to a wave while [6; 6; 6] infected six clusters only. Similarly, nine initial cases could seed either zero local cluster or a few, depending on how they were distributed over the three days. To some extent, this variation might be the effect of the roundoff routine we are using – we shall clarify this in the next Section.

As examples of the converse situation where a stable region is seeded with a very high caseload, we consider two scenarios in Fig. 8. In the top panel, we take the default parameter values and seed it with a vector of external cases as above, but we choose this vector to be [120; 120; 120]. In the bottom panel, we start from the parameter values of Fig. 7 (default plus $n_S = 1000$) and seed it with the 8 initial clusters of Figs. 4-6. One hundred days into the epidemic however, we slash n_S to zero to bring all parameters to their default values.

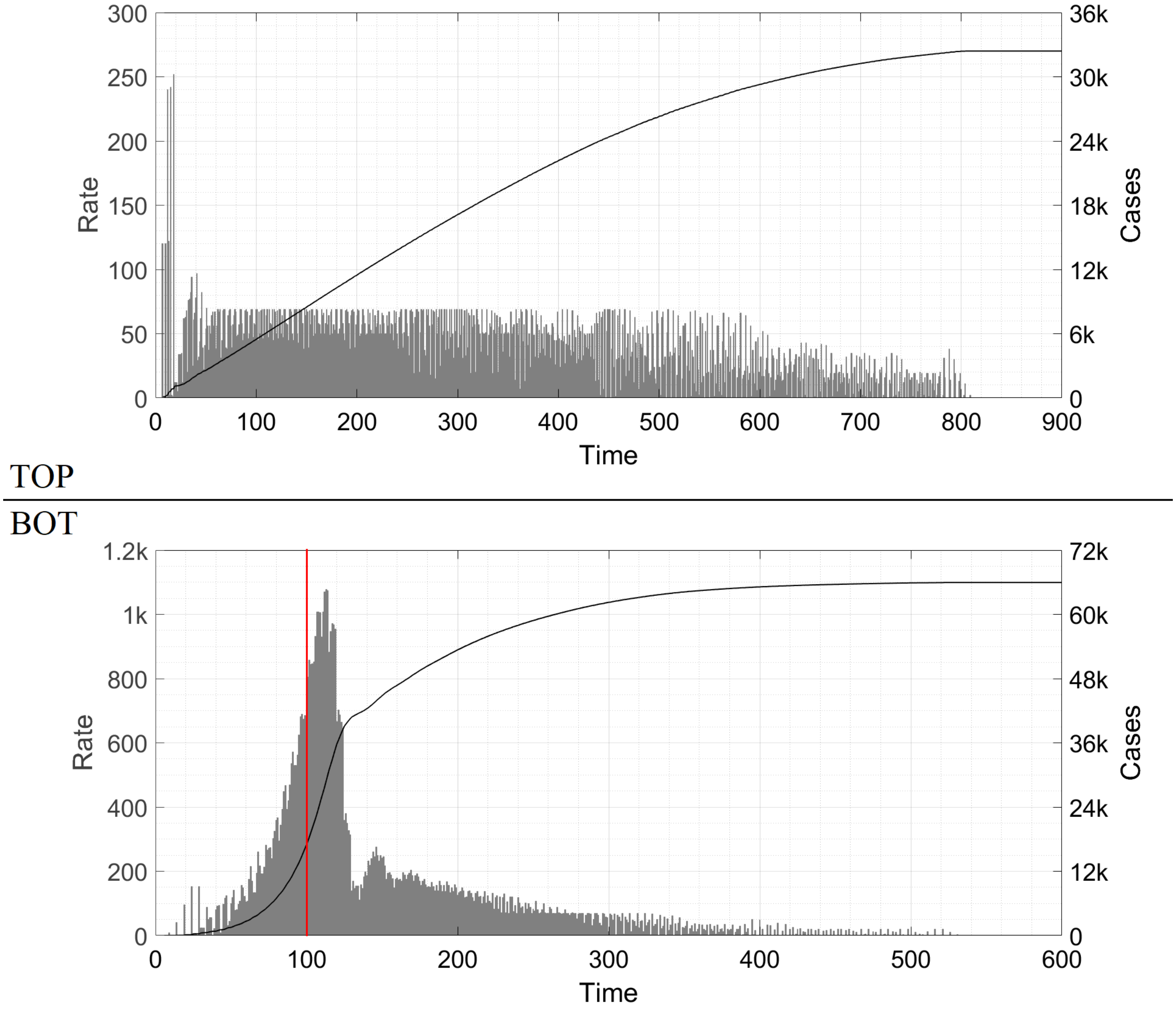

Figure 8 : *Time traces of the epidemic to demonstrate additional seeding-related effects. In the bottom panel the red line at day #100 indicates the abrupt reduction of* ns *from 1000 to 0. The symbol 'k' denotes thousand.*

In the top panel, the cumulative caseload is higher than in Fig. 4 but the nature of the solution remains qualitatively unchanged. In the bottom panel, the cases decrease rapidly after the brakes are hit and the epidemic plateaus before eventually petering out.

§14. Critical mass effect. Figure 7 shows yet another phenomenon which is absent in classical epidemiological models. In these models, when the parameters are chosen to generate an instability (for example, in the DDE model [9] if the spreading rate m_0 is taken above the critical value), even the smallest non-zero IC is sufficient to set off a wave of disease. Since the early growth is exponential, it hardly matters whether the IC features 100 cases or 1 case or 0.01 case (this latter being perfectly legitimate in a continuous model) – a small seeding can at most account for a delay in the peak by a couple of doubling times. Here however we see a huge difference. The parameters in Fig. 7 are chosen to lie in the highly unstable region, as the bottom panel demonstrates. Even so, it is possible for the outbreak to stop at just the seeding cases (top panel) or at the seeding cases plus a handful of clusters (middle panel).

We have already mentioned an anomaly regarding the ICs [4; 4; 4] vis-a-vis [6; 6; 6]. We believe that this is due to errors accumulated in the roundoff process – a lucky combination of values might infect a whole cluster and terminate the run while an unlucky combination might keep alive a fractional cluster which eventually adds up and perpetuates the epidemic. The presence of a discrepancy warrants a detailed analysis of the IC to verify that the effects shown in Fig. 7 are not spurious. We perform this calculation in the Appendix; the result of this procedure is that the effect is genuine.

Thus, even when a city is operating in the unstable region of parameter space, it needs a small but finite minimum number of initial cases to set the wave off. By analogy with nuclear reaction theory where a finite minimum quantity of fissile material is required to initiate the chain reaction, we call the seeding threshold the **critical mass**. From the viewpoint of epidemic management, the existence of critical mass is extremely dangerous. Seeded below this minimum, a city which is actually unstable will falsely behave like a stable or neutral region, conveying a deceptive impression that the disease is under control when it is actually a disaster waiting to happen.

Figure 8 indicates that the converse of the critical mass phenomenon is not true – when the parameters are in the neutral region, even a huge seeding cannot turn it unstable. We have checked this result for other parameter and seeding combinations as well and found it to be general. This is in agreement with the predictions of classical models where stability is independent of initial condition. It is a positive outcome from the viewpoint of epidemic management, since it implies that once tight controls are re-established following a surge, the presence of an existing huge number of cases will not cause the epidemic to propagate by itself like a perpetual motion machine. India and Taiwan are excellent demonstrators of this.

While the general conclusions of Fig. 8 are robust, we do not set too great a store by Fig. 8-bot as an indicator of lockdown dynamics, especially if the lockdown is hard. This is because in a lockdown the clusters are forcibly broken up – groups of friends do not have the option to meet and socialize at entertainment venues. While case counts are expected to remain flat or even increase for a few days following lockdown on account of the serial interval and the household transmissions, the sharp increase seen in Fig. 8-bot between days #100 and #110 is less realistic as it arises from mechanically implementing the cluster sequence on all clusters seeded upto day #99. Similarly, the precipitous decline after day #110 and the dip near day #130 are possibly numerical artefacts. The rather slow decrease in case rate from day #150 onwards is however a most robust prediction – when case levels have fallen after a surge, we expect entertainment venues to be reopened which will resume intra-cluster transmission. In this regime, a persistent case level might really be prevalent for a long time, and vaccination will be the best technique to drive it down.

---- o ----

DISCUSSION

§15. Qualitative explanation of the wave in Delhi. Since we took this city as a case study of the limitations of classical models, we now outline how CST model is able to account for the most counter-intuitive features of the wave in Delhi. The low case counts despite open public transport and entertainment venues can be explained as stable or neutral operation with high intra-cluster but low inter-cluster transmission. This is a marked contrast from the herd immunity explanation of classical models. The suddenness of the wave with no early warning can be attributed to the critical mass effect. As we have seen, this effect can result in suppression of an instability when the absolute case counts are sufficiently small. We conjecture that this is just what happened in Delhi – an increase in SEC events like weddings and parties activated the cryptogenic instability some time in January or February itself (see later for evidence in support of this conjecture). However, by then, the absolute case counts were so low that in the first few weeks after breaching the stability threshold, the wave was not initiated. Unaware of the threat, social gatherings continued picking up pace and the operating point drifted farther and farther into the unstable zone.

The critical mass was eventually attained via influx from Maharashtra. Delhi and Mumbai are the pair of cities in India with the maximum number of connecting flights – in fact, the air corridor is the fifth busiest in the world. By mid-March, Maharashtra was already seeing high number of cases; passengers carrying virus almost certainly arrived into Delhi and spread the disease to the local community. Only now was Delhi's true nature of operation revealed, and with disastrous consequences. The critical mass effect can also explain why the two 2020 waves shown in Fig. 1 were in line with classical expectations. In both instances, when the waves started, the absolute case counts were much higher – about ten times greater than before the 2021 wave. These numbers exceeded the critical mass and hence, as soon as there was an increase in mobility across a stability threshold, the case counts started going up. The rate of increase was slow, as expected for a minor instability, and NPI served to flatten and thereafter bring down the curves.

A similar phenomenon occurred during the second wave in Maharashtra, where we recall that R of the second wave was lower than in Delhi despite being driven by the same transmissible double mutant strain. Here also, the absolute and per-capita counts of cases were much higher – in January 2021, Pune for example had a minimum daily case rate about double of Delhi despite having one-sixth the population. Such trends held true in the rest of Maharashtra as well and the state remained above critical mass throughout. Hence, cases started increasing as soon as the cryptogenic instability was activated, which was in mid-February. Consistent with expectation, the increase was slow. Once an increase was manifest, it automatically cast a damper on social functions and prevented the reproduction number, though above unity, from climbing even higher. The cultural similarities across big cities in India motivate the conjecture that Delhi's stability transition had also occurred at around this time itself, as mentioned above.

Hence, cryptogenic instability and critical mass effects are capable of explaining all the paradoxical phenomena associated with the 2021 wave in Delhi, including the points mentioned in our discussion of the analysis in Ref. [24].

§16. Limitations and classical limit. Here we discuss the assumptions in the model. A lot of the assumptions stem from our desire to keep the model deterministic even though it has probabilities at its core. Determinism is necessary for computational tractability – a typical run with high caseload like Fig. 7-bot takes about five minutes on a laptop computer. Since not everyone (including ourselves) has access to a high-performance computing cluster, we have tried, at least in this Article, to create a basic model which is highly accurate but remains solvable on a commonly available computer. We now show that almost all of the model assumptions can be removed or relaxed with increased computational capacity. Some obvious assumptions here are those of constant household and cluster size. A more advanced formulation of the model can incorporate a distribution of household and cluster sizes and can also account for the fact that more than one household member is socially active (if two household members are active for example, then their respective clusters effectively get merged into a double-sized cluster).

Another assumption here is the constant cluster sequence of [1; 3; 6; 7; 5; 1], which remains valid even if multiple members of the same cluster are infected simultaneously. A more accurate implementation of the intra-cluster dynamics can be achieved by prescribing an interaction pattern among cluster members, prescribing the probability of transmission with each interaction and simulating the epidemic's evolution. This however makes the model non-deterministic. As an approximate "average" value, the cluster sequence is adequate, since it incorporates the same logic at the cluster level as the DDE model [9] does at the population level. Accounting for multiple seeding (i.e. more than one initial case in a cluster) will make the cluster get infected faster and thus bring forward a few cases by a few days; it will not affect the total count however. Moreover, it turns out that the basic phenomena we see here are independent of the specific choice of cluster sequence. To demonstrate this, we consider a run with all parameter values same as in Fig. 4 but the cluster sequence replaced by [1; 4; 13; 6]. This corresponds to a more transmissible strain of virus with an R_0 of approximately 4. The result is shown in Fig. 9 below.

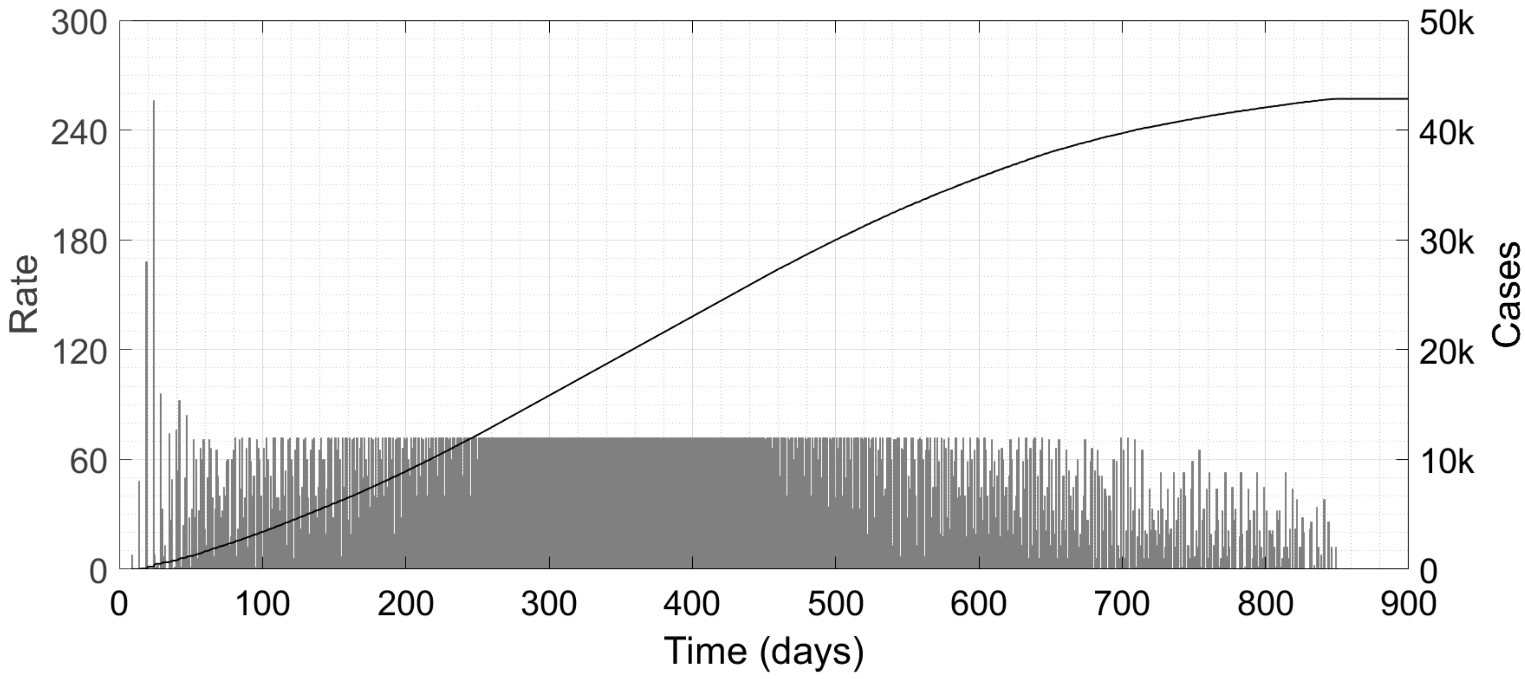

Figure 9 : *Time trace of the epidemic with a different cluster sequence, corresponding to faster transmission within the cluster.*

We can again see the constant rate trajectory from Fig. 4 and Fig. 5-mid. The cumulative caseload is greater than in Fig. 4, which is consistent since the new cluster sequence assumes faster transmission (and hence a greater number of at large cases at any time) than the old one. Nearly identical solutions with two widely disparate cluster sequences convinces us that the specific details of the sequence are not too important in determining the solution. We have also checked (not shown here) that the plateau of Fig. 4 again undergoes the cryptogenic instability as n_S is increased; the value of n_S required for a large wave is lower than in Fig. 6, and becomes lower still if we increase m_S and P_U to accommodate a more transmissible viral strain. We dispense with a repetition of the plots for these situations since they are qualitatively the same as Figs. 4-8; only the parameter values for stability transitions etc are different.

A few other approximations deserve a quick mention. The derivation of (7) and (8) contains a caveat mentioned in the Appendix while the introduction of the parameter $k_{\max}$ has been discussed in §7; these do not generate significant error. Similarly, the errors arising from roundoff (for example the overly smooth plateau of §9) are also small.

A fully agent-based implementation of this model – in which there are only lattice sites and probabilities – can do away with almost all the assumptions mentioned above. The beginnings of such an approach can be seen in the validation of the critical mass effect in the Appendix. Such an approach will enable us to answer questions like "Given there are 100 cases today, what is the probability that there will be 1000 cases a month later ?" by calculating the probability of occurrence of every possible path from 100 to 1000 over 30 days (a daunting task by any standards). Such calculations can also yield best and worst case scenarios of the disease evolution corresponding to any given level of NPI. As we have already mentioned in §1, the accuracy of an agent-based model depends significantly on the network structure assumed by the modellers. Our work here shows that social heterogeneity should be at the core of any realistic network structure.

Despite the approximations, our model is capable of achieving excellent agreement at the qualitative level, as the analysis of Delhi shows. For a meaningful demonstration of quantitative agreement, for example with the Delhi data, we need to be able to know, from a source external to the COVID-19 trajectories themselves, the values of the social interaction parameters such as cluster size, n_S and n_U. If these values become known, then the transmission-related quantities such as m_S and P_U can be estimated from a fit of the data. We also need to be able to account for contact tracing, an intervention used very widely in Delhi (and in most other places worldwide), which through preferential isolation of suspected cases can result in significantly lower levels of disease at higher levels of mobility.

In fact, an analysis of contact tracing can provide a key to the generation of accurate estimates for the parameters in our model. As already mentioned, most of these parameters are related to interaction, and the best source of obtaining interaction data is from the contact tracing authorities of a particular city or town. Currently we do not have access to any such data (which is invariably privacy-protected and never reported in public databases or research works), and a fit will involve too many free parameters. Hence we leave the data fit for a future study. We do expect however that, just as the bell-shaped epidemic curve is a generic solution of the S-I-R model whatever the parameter values, the plateau solution and the phenomena of cryptogenic instability and critical mass effect are generic solutions of the present model and are not dependent on the specific parameter values which we have used here.

To further boost credibility of our model, we demonstrate that it reduces to a classical model in the appropriate limit. The details of this procedure are in the Appendix but the bottom line is that, under certain near-trivializing assumptions, our model can be transformed into the DDE [9].

§17. Conclusion. In this Article we have proposed the cluster seeding and transmission (CST) model for the spread of COVID-19 which takes into account the heterogeneity in human interaction patterns. This model has several features which are absent from conventional epidemiological models. The first is the presence of the plateau solution, or constant daily case rate, as a generic feature. Secondly, we have found that socializing events can destabilize the plateau into a gigantic wave even though the increase in population-averaged contact rate remains minuscule. We have called this the cryptogenic instability. Finally we have observed that the initial conditions play a very important role as well. In particular, even if the parameters are chosen to ensure uncontained growth of the epidemic, a sufficiently small IC can result in the growth not being manifest. We have called this the critical mass effect. We have presented the city of Delhi as a case study where the classical epidemic models failed at the qualitative level but CST model can explain the counter-intuitive observations.

In future work, we shall analyse how the phenomena outlined here lead to quantitative explanation of the case trajectories in Delhi as well as other cities and countries. We hope to secure interaction data from contact tracing agencies to generate meaningful fits to real-world data. Adding measures like contact tracing and vaccination to the basic model, we shall also attempt to calculate the effects of various kinds of interventions, both pharmaceutical and non-pharmaceutical. Such additions will enable our analysis to be used for making public health policy decisions, which is the long-term goal of several mathematical infectious disease models. While COVID-19 remains our immediate focus, the concepts we have presented here are valid for any infectious disease. For example, the SARS-1 outbreak of 2002-3 became contained because it had lower values of P_U and m_S as well as reduced intra-cluster transmission as compared to COVID-19. We hope that our model and findings may have utility in the management of future epidemics as well.

---- o ---- o ---- o ----   ---- o ---- o ---- o ----

ACKNOWLEDGEMENTS

We would like to thank the two anonymous Reviewers of AIP Chaos, whose comments have resulted in significant improvement to the quality of this manuscript.

DECLARATIONS

We have NO conflict of interest to declare.

We have NOT been funded by any agency, public or private, for this study.

DATA AVAILABILITY

All data used in this study are publicly available. All codes will be shared on request, either for verification purposes or for novel uses; please contact SHAYAK at sb2344@cornell.edu or shayak.2015@iitkalumni.org for the same.

REFERENCES

[1] WO Kermack and AG McKendrick, "A Contribution to the mathematical theory of epidemics," **Proceedings of the Royal Society A** 115 (772), 700-721 (1927)
[2] R Ranjan, A Sharma, and MK Verma, "Characterization of the Second Wave of COVID-19 in India," **MedRxiv Article** (2021). Available at https://www.medrxiv.org/content/10.1101/2021.04.17.21255665v2
[3] R Li, S Pei, B Chen et. al., "Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2)," **Science** 368 (6490), 489–493 (2020)
[4] G Giordano, F Blanchini, R Bruno et. al., Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy, **Nature Medicine** 26 (6), 855-860 (2020)
[5] RM Anderson, JAP Heesterbeek, D Klinkenberg and TD Hollingsworth, "How will country-based mitigation measures influence the course of the COVID-19 epidemic?," **The Lancet** 395 (10228) 931–934 (2020)
[6] J Dehning, J Zierenberg, FP Spitzner et. al., "Inferring change points in the spread of COVID-19 reveals the effectiveness of interventions," **Science** 369 (6500), eabb9789 (2020)
[7] K Prem, Y Liu, TW Russell et. al., "The Effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study," **The Lancet Public Health** 5 (5) e261–e270 (2020)
[8] M Agrawal, M Kanitkar and M Vidyasagar, "SUTRA: An Approach to Modelling Pandemics with Asymptomatic Patients, and Applications to COVID-19," **Arxiv Article** (2021). Available at https://arxiv.org/abs/2101.09158
[9] B Shayak and MM Sharma, "A New Approach to the Dynamic Modeling of an Infectious Disease," **Mathematical Modelling of Natural Phenomena** (2021). Available at https://www.mmnp-journal.org/component/article?access=doi&doi=10.1051/mmnp/2021026
[10] CC Kerr, RM Stuart, D Mistry et. al., "Covasim : an agent-based model of COVID-19 dynamics and interventions," **PLOS Computational Biology** 17 (7), e1009149 (2021)
[11] SR Serrao, S Deng, P Priyanka et. al., "Requirements for the containment of COVID-19 disease outbreaks through periodic testing, isolation and quarantine," **MedRxiv Article** (2020). Available at https://www.medrxiv.org/content/10.1101/2020.10.21.20217331v1
[12] NC Grassly, E Pons-Salort, EPK Parker et. al., "Comparison of molecular testing strategies for COVID-19 control: a mathematical modelling study," **The Lancet Infectious Diseases** 20 (12), 1381–1389 (2020)
[13] J Hellewell, S Abbott, A Gimma et. al., "Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts," **The Lancet Global Health** 8 (4), e488–e496 (2020)
[14] S Agrawal, S Bhandari, A Bhattacharjee et. al., "City-scale agent-based simulators for the study of non-pharmaceutical interventions in the context of the COVID-19 pandemic," **Arxiv Article** (2020), Available at https://arxiv.org/abs/2008.04849
[15] G Pescarmona, P Terna, A Acquadro et. al., "An Agent-based model of COVID-19 diffusion to plan and evaluate intervention policies," **Arxiv Article** (2021). Available at https://arxiv.org/abs/2108.08885
[16] JM Cashore et. al., "COVID-19 mathematical modeling for Cornell's fall semester," (2020). Available at https://cpb-us-w2.wpmucdn.com/sites.coecis.cornell.edu/dist/3/341/files/2020/10/COVID_19_Modeling_Jun15-VD.pdf
[17] D Adak, A Majumder and N Bairagi, "Mathematical perspective of COVID-19 pandemic : disease extinction criteria in deterministic and stochastic models," **MedRxiv Article** (2020). Available at https://www.medrxiv.org/content/10.1101/2020.10.12.20211201v1

[18] Y Gu, “Estimating true infections : a simple heuristic to measure implied infection fatality rate,” (2020). Available at https://covid19-projections.com/estimating-true-infections/
[19] F Ball, D Morrison and G Scalia-Tomba, “Epidemics with two levels of mixing,” **The Annals of Applied Probability** 7 (1), 46-89 (1997)
[20] C Moore and MEJ Newman, “Epidemics and percolation in small-world networks,” **Physical Review E** 61 (5), 5678-5683 (2000)
[21] M Kuperman and G Abramson, “Small world effect in an epidemiological model,” **Physical Review Letters** 86 (13), 2909-2912 (2001)
[22] MEJ Newman, “Spread of epidemic disease on networks,” **Physical Review E** 66 (1), 1-11 (2002)
[23] JS Juul and SH Strogatz, “Descendant distributions for the impact of mutant contagion on networks,” **Arxiv Article** (2019). Available at https://arxiv.org/abs/1910.00655
[24] MS Dhar, R Marwal, VS Radhakrishnan et. al., “Genomic characterization and epidemiology of an emerging SARS-CoV-2 variant in Delhi, India,” MedRxiv Article (2021). Available at https://www.medrxiv.org/content/10.1101/2021.06.02.21258076v1
[25] S Thurner, P Klimek and R Hanel, “A network-based explanation of why most COVID-19 infection curves are linear,” **Proceedings of the National Academy of Sciences USA (PNAS)** 117 (37), 22684–22689 (2020)
[26] L Kusmierz and T Toyiozumi, “Infection curves on small-world networks are linear only in the vicinity of the critical point,” **Proceedings of the National Academy of Sciences USA (PNAS)** 118 (10), e2024297118 (2021)
[27] AV Tkachenko, S Maslov, T Wang et. al., “Stochastic social behaviour coupled to COVID-19 dynamics leads to waves, plateaus and an endemic state,” **MedRxiv Article** (2021). Available at https://www.medrxiv.org/content/10.1101/2021.01.28.21250701v3
[28] BF Nielsen, L Simonsen and K Sneppen, “COVID-19 superspreading suggests mitigation by social network modulation,” **Physical Review Letters** 126 (11), 118301 (2021)
[29] S Manrubia and DH Zanette, “Individual risk aversion responses tune epidemics to critical transmissibility (R=1),” **Arxiv Article** (2021). Available at https://arxiv.org/abs/2105.10572
[30] T Kar, S Sarkar, S Chowdhury et. al., “Anticipating the novel coronavirus disease (COVID-19) pandemic,” **Frontiers in Public Health** 8, 569669 (2020)
[31] F Dablander, JAP Heesterbeek, D Borsboom and JM Drake, “Overlapping time-scales obscure early warning scales of the second COVID-19 wave,” **MedRxiv Article** (2021). Available at https://www.medrxiv.org/content/10.1101/2021.07.27.21261226v3
[32] B Hejazi, A Schlenczek, B Thiede, G Bagheri and E Bodenschatz, “Aerosol transport measurements and assessment of risk from infectious aerosols : a case study of two German cash-and-carry hardware/DIY stores,” **Arxiv Article** (2021). Available at https://arxiv.org/abs/2105.10357
[33] H Nishiura, NM Linton and AR Akhmetzhanov, “Serial interval of novel coronavirus (COVID-19) infections,” **International Journal of Infectious Diseases** 93, 284–286 (2020)
[34] B Rai, A Shukla and LK Dwivedi, “Estimates of serial interval for COVID-19 : a systematic review and meta-analysis,” **Clinincal Epidemiology and Global Health** 9 (1), 157–161 (2021)
[35] ML Childs, MP Cain, D Kirk et. al., “The Impact of long term non-pharmaceutical interventions on COVID-19 epidemic dynamics and control,” **MedRxiv Article** (2020). Available at https://www.medrxiv.org/content/10.1101/2020.05.03.20089078v1
[36] M D’Arienzo and A Coniglio, “Assessment of the SARS-CoV-2 basic reproduction number, R0, based on the early phase of COVID-19 outbreak in Italy,” **Biosafety and Health** 2 (2), 57–59 (2020)
[37] NG Davies, S Abbott, RC Barnard et. al., “Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England,” **Science** 372 (6538) eabg3055 (2021)
[38] ED Laing, NJ Epsi, SA Richard and EC Samuels, “SARS-CoV-2 antibodies remain detectable 12 months after infection and antibody magnitude is associated with age and COVID-19 severity,” **MedRxiv Article** (2021). Available at https://www.medrxiv.org/content/10.1101/2021.04.27.21256207v1
[39] KW Cohen, SL Linderman, Z Moodle et. al., “Longitudinal analysis shows durable and broad immune memory after SARS-CoV-2 infection with persisting antibody responses and memory B and T cells,” **MedRxiv Article** (2021). Available at https://www.medrxiv.org/content/10.1101/2021.04.19.21255739v1
[40] J Mossong, P Hens, M Jit et. al., “Social contacts and mixing patterns relevant to the spread of infectious diseases,” **PLOS Medicine** 5 (3), 381–391 (2008)
[41] K Backlawski and GC Rota, "*An Introduction to Probability and Random Processes*" (1979)

APPENDIX

§A1. Derivation of the probability expressions. Here we derive (2) in detail. The problem statement is : given that there are total N_C clusters and that β people have been exposed to the virus, what is the probability that they belong to b clusters $(1 \leq b \leq \beta)$? With one caveat, this problem is identical to the following ‘toy’ problem : there are N_C boxes and β balls; if any ball can go into any box, what is the probability that exactly b boxes contain at least one ball ? The caveat is that in the balls and boxes problem, a single box might contain all the balls while in the cluster problem, one cluster cannot contain more than 24 people.

We ignore the caveat for the following reason : in a typical situation, $N_C >> \beta, b$ (during a typical run, a b is approximately 40 at the height of the epidemic wave, about two orders of magnitude less than N_C which is 4200) and there will be a very large probability that β people will belong to β or nearly β clusters. The spurious occurrences which we will pick up by ignoring the caveat (β people belonging to less than $\lceil \beta / 24 \rceil$ clusters) will be extremely improbable anyway and the convenience gained will amply recompense the losses incurred. The caveat surmounted, the balls and boxes problem happens to be solvable in closed form. The numerator of the probability must be the number of ways of putting β balls into exactly b boxes ($b \leq \beta$), while the denominator must include all possible ways of putting the balls into the boxes. For the numerator, we first select b boxes among the N_C ones, which can be done in ${}^{N_c}\mathrm{C}_b$ ways. Given the box selection, we must find the number of ways to express β as a sum of b natural numbers, including different orderings – this is the restricted composition function $\tilde{p}(b,\beta)$. The composition and not the partition function $p(b,\beta)$ is relevant for the following reason : suppose we have 10 boxes in total and want to distribute 5 balls among 3 boxes, and suppose we have already selected the boxes number 2, 4 and 7. Then, it makes a big difference whether there are three balls in box 4 and one each in boxes 2 and 7 or three balls in box 2 and one each in boxes 4 and 7. The composition function $\tilde{p}(3,5)$ treats 3+1+1 as different from 1+3+1 and accounts for this difference while the partition $p(3,5)$ treats the two as identical and fails to account for it.

The composition $\tilde{p}(b,\beta)$ actually has a simple analytical formula unlike the partition. We can view a composition of β into b parts as introducing $b-1$ slats in between β beads arranged in a row on an abacus. So long as there is maximum one slat between two beads and there are no slats to the left or right of the row as a whole, the slats split the beads into exactly b parts – each different arrangement of slats corresponds to a different way of writing β as a sum of b numbers. In total, there are $\beta-1$ gaps between the beads which need to be filled by $b-1$ slats; the number of ways of doing this is ${}^{\beta-1}\mathrm{C}_{b-1}$. Thus, $\tilde{p}(b,\beta) = {}^{\beta-1}\mathrm{C}_{b-1}$.

To calculate the denominator of the balls and boxes probability, we imagine the N_C boxes lying in a row and focus on the walls separating one box from the next. The external walls of the leftmost and rightmost box must remain intact; the remaining N_C-1 walls plus the β balls can be arranged in a line in any order whatsoever. For this, among the total $N_C-1+\beta$ positions which can be filled by either ball or wall, select β to fill with balls; the remainder automatically get filled with walls. The number of ways of doing this selection is ${}^{N_c-1+\beta}\mathrm{C}_{\beta}$. Putting all this together, and switching back from balls and boxes to viral recipients and clusters, the probability that β recipients belong to b clusters is given by (2).

§A2. Validation of the critical mass effect. Our intention here is to demonstrate that the critical mass effect is genuine and not an artefact of the numerical roundoff process. There are two steps where rounding off takes place – once when calculating γ as rounded off kP_U and again when calculating the expectation value $E(\Delta z_i)$. For really small numbers of cases, the concepts of roundoff or expectation value do not have too much meaning. A more relevant question is : given that there are α cases at large today, what is the probability that the virus is introduced into exactly 0, 1, 2 etc new clusters ? Since we are dealing with the start of the outbreak, we assume that all clusters are susceptible.

The bulk of the calculational framework we have already developed in §6; a few extras needed to be taken care of. First is the probability that there are zero infected clusters – this eventuality was not relevant for calculating the expectation value as it would have had a null contribution. At SEC events, each case transmits to $m_S = 2$ people by definition, so zero transmission can occur if and only if zero cases are present at the events. The probability of this happening is

$$P_{\text{SEC}=0} = \frac{{}^{N_1-\alpha}\mathrm{C}_{n_S}}{{}^{N_1}\mathrm{C}_{n_S}} \quad . \tag{A1}$$

The calculation for $P_{\text{UCT}=0}$ has one difference; while successful transmission to two people at SEC events is a certainty, transmission at UCT events is probabilistic with a chance of P_U. From now onwards, we treat P_U as a genuine probability and not as an averaged quantity equivalent to m_S, i.e. we assume that at each UCT event, a case transmits the disease to exactly one person with probability P_U and to zero person otherwise. Consequently, for $P_{\text{UCT}=0}$, we must take into account not only the situation where there are no at large cases attending UCT events but also where there are k such cases and just none of them happen to transmit. The probability of there being k cases participating in UCT events has already been calculated in §6; the probability that none transmit is $(1-P_U)^k$. The probability of $k = 0$ is the exact equivalent of (A1); taking this term and adding all the terms for k going from 1 to α yields

$$P_{\text{UCT}=0} = \frac{{}^{N_1-\alpha}\mathrm{C}_{n_U}}{{}^{N_1}\mathrm{C}_{n_U}} + \sum_{k=1}^{\alpha} \frac{{}^{\alpha}\mathrm{C}_k \; {}^{N_1-\alpha}\mathrm{C}_{n_U-k}}{{}^{N_1}\mathrm{C}_{n_U}} \left(1-P_U\right)^k \quad . \tag{A2}$$

We now focus on infection of non-zero numbers of clusters.

The probability that SEC events infect exactly b clusters has already been calculated as part of (6); the expression is

$$\begin{aligned} P_{\text{SEC}=b} &= \sum_k P(k)P(b\,|\,k) \\ &= \sum_{k=1}^{\alpha} \frac{{}^{\alpha}\mathrm{C}_k \; {}^{N_1-\alpha}\mathrm{C}_{n_S-k}}{{}^{N_1}\mathrm{C}_{n_S}} \left(\frac{{}^{N_C}\mathrm{C}_b \; {}^{2k-1}\mathrm{C}_{b-1}}{{}^{N_C-1+2k}\mathrm{C}_{2k}} \right) \quad . \end{aligned} \tag{A3}$$

The probability that UCT events infect b clusters consists of three sub-probabilities, • that there are k cases at large, • that these k cases infect j new people with j going from 1 to k, and • that these j infectees belong to b clusters. The first and third of these probabilities have already been calculated in §6; the second is identical to the probability that k tosses of a biased coin with probability P_U of heads result in j heads. This is ${}^{k}\mathrm{C}_j \, P_u^j \left(1-P_U\right)^{k-j}$; to obtain $P_{\text{UCT}=b}$ we must (as usual) multiply the sub-probabilities and sum over both k and j getting

$$P_{\text{UCT}=b} = \sum_{k=1}^{\alpha} \frac{{}^{\alpha}\mathrm{C}_k \; {}^{N_1-\alpha}\mathrm{C}_{n_S-k}}{{}^{N_1}\mathrm{C}_{n_S}} \left[\sum_{j=1}^{k} {}^{k}\mathrm{C}_j \, P_U^j \left(1-P_U\right)^{k-j} \left(\frac{{}^{N_C}\mathrm{C}_b \; {}^{j-1}\mathrm{C}_{b-1}}{{}^{N_C-1+j}\mathrm{C}_j} \right) \right] \quad . \tag{A4}$$

It is now a simple matter to compute the probability that α at large cases infect a total of b clusters. For b taking the values 0, 1 and 2 we have

$$P_{b=0} = P_{\text{SEC}=0} P_{\text{UCT}=0} \quad , \tag{A5a}$$

$$P_{b=1} = P_{\text{SEC}=1} P_{\text{UCT}=0} + P_{\text{SEC}=0} P_{\text{UCT}=1} \quad , \tag{A5b}$$

$$P_{b=2} = P_{\text{SEC}=2} P_{\text{UCT}=0} + P_{\text{SEC}=1} P_{\text{UCT}=1} + P_{\text{SEC}=0} P_{\text{UCT}=2} \quad . \tag{A5c}$$

In Fig. 10, we plot these probabilities as a function of α for the latter taking the values from 1 to 100. We use the parameter values of Fig. 7. We also plot $1 - P_{b=0} - P_{b=1} - P_{b=2}$, which is the probability that three or more clusters are infected.

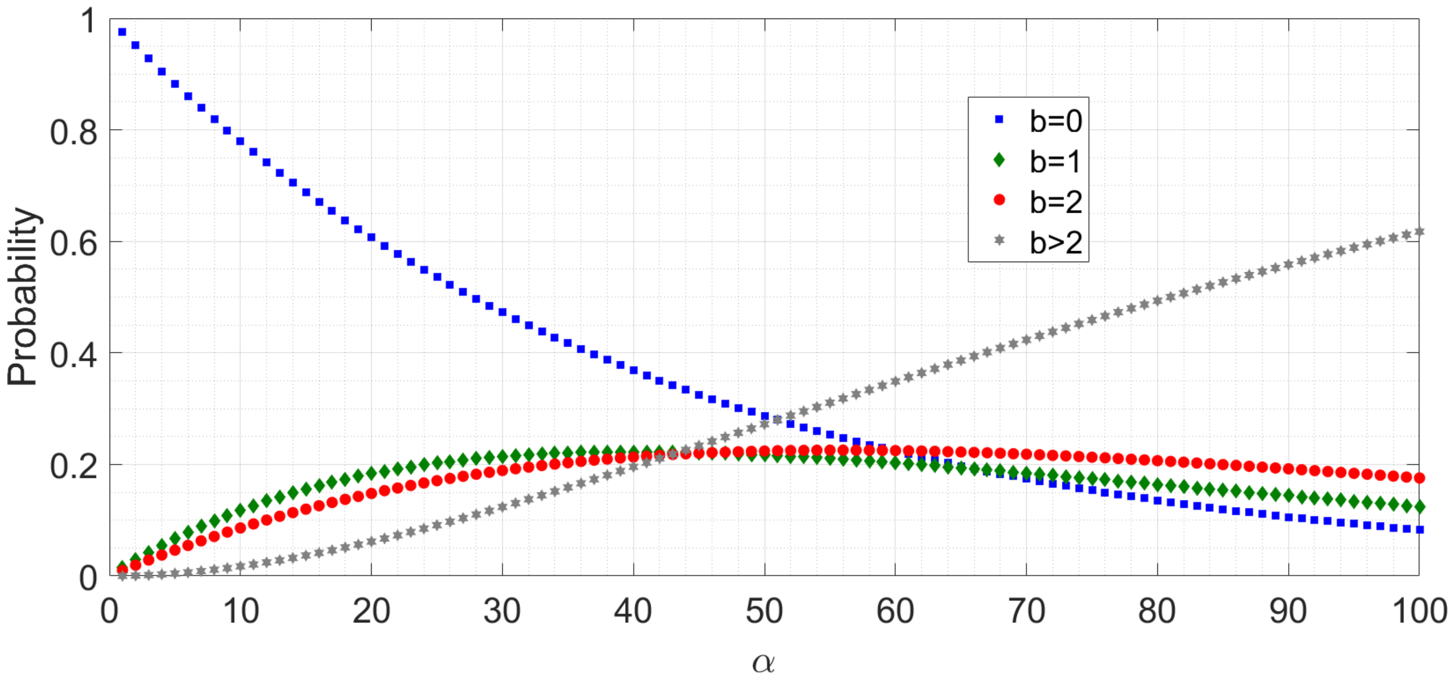

Figure 10 : *Probability of different number of clusters' getting infected on any particular day, given the number of at large cases on that day.*

This Figure shows that for low α, the likelihood of zero cluster's being infected is very close to unity. As expected, this likelihood decreases as α increases but even at $\alpha = 28$ the probability of no new infection is 1/2. Only at $\alpha = 51$ does zero infection cease to be the most probable outcome. Thus, at low α, corresponding to a small external case influx or 1-2 infected clusters, it is indeed quite possible that the infection will not advance further in the population. This provides justification for the findings of Fig. 7.

§A3. Reduction to the DDE model. To obtain the DDE model [9] from CST model, we must first incorporate homogeneous mixing (the unstated or at best understated pillar of classical models). For this we consider the situation where households are scrapped (so that $N = N_1$) and clusters are reduced to size unity ($N_C = N$, $y_i = z_i$ and $\Delta y_i = \Delta z_i$ for all i). In this limit, UCT and SEC modes of transmission are equivalent; let us merge them into the SEC category. m_S remains the number of people to whom one case spreads the disease in a day if they are all susceptible. To incorporate continuous mixing (another classical assumption) we let $n_S = N$ i.e. we allow the entire population to mix every day. Under these assumptions let us run through the algorithmic procedure outlined in §7.

In a classical model, we do not need rounding off so we discard the subroutine roundoff. In the main routine, everything upto the primary loop over i remains unchanged. In the primary loop we let go of $k_{\max}$. Then, (6) as it stands no longer has meaning. Since everyone participates in the 'SEC' events, k can take only the value α – the first summation and the expression for $P(k)$ both vanish. The number of people who receive the infective dose of virus remains $\beta = m_s\alpha$. By definition these β people belong to exactly β 'clusters' so the second summation in (6) is redundant as well. What remains is the third summation i.e. the calculation of the expected number of susceptible clusters which are seeded (or equivalently the expected number of susceptible persons who are infected).

The problem we have on hand is, given that there are $N - y_i$ people who are susceptible and y_i people who are immune, what is the probability $P(j)$ that among β randomly selected people who receive an infective dose, exactly j are susceptible ? Then the expectation value can be calculated as $E = \Sigma_j jP(j)$ as previously, with j running from 1 to β. The fully accurate expression for $P(j)$ is

$$P(j) = \frac{{}^{N-y_i}\mathrm{C}_j \; {}^{y_i}\mathrm{C}_{\beta-j}}{{}^{N}\mathrm{C}_{\beta}} \quad , \tag{A6}$$

as we have already calculated several times in different contexts. However, if $\beta << y_i$, $N–y_i$, i.e. the daily new cases are very low compared to the susceptible and immune populations (a very reasonable assumption even in the worst of worst-hit areas unless we are extremely close to the beginning of the epidemic), then we can assume that each individual person is susceptible with probability $p = 1 - y_i/N$ and immune with probability $1-p$. (An equivalent problem is that a box contains y_i blue balls and $N-y_i$ green balls; if β balls are drawn at random without replacement, then what is the probability that j balls are green ? In the limit y_i, $N–y_i >> \beta$, we can replace this by the situation where the balls are drawn *with* replacement.) With this assumption, $P(j)$ becomes

$$P(j) = {}^{\beta}\mathrm{C}_j\, p^j (1-p)^{\beta - j} \quad . \tag{A7}$$

The expectation value of a binomial distribution is a standard formula in probability theory [41]; in our problem it evaluates to $\beta p = \beta(1 - y_i/N)$. Thus, instead of the expression (6) we now use the value $E_S = \beta(1 - y_i/N)$.

Since UCT events have been merged with SEC events we skip the steps pertaining to the calculation of E_S. Implementation of the cluster sequence remains as is with [1; 3; 6; 7; 5; 1] being replaced by singleton vector [1] (not Reference 1). So, putting everything together and using the values of α and β, we have

$$\Delta y_{i+5} = m_0 \left(1 - \frac{y_i}{N}\right)\left(\Delta y_{i-1} + \Delta y_{i-2} + \Delta y_{i-3}\right) \quad , \tag{A8}$$

where, since SEC and UCT events have lost their separate identities, m_0 is more appropriate than m_S. We now want to express this as a continuous time system or flow rather than a map, so let the continuous function $y(t)$ denote the cumulative count of corona cases as a function of time. Δy_i is the discrete equivalent of $\mathrm{d}y/\mathrm{d}t$; the big bracket in the right hand side of (A8) above is the number of new cases which have cropped up over the last three days, which is $y(t) - y(t-3)$. Using this we have

$$\left.\frac{\mathrm{d}y}{\mathrm{d}t}\right|_{t+5} = m_S \left(1 - \frac{y}{N}\right)\left[y - y(t-3)\right] \quad \text{, or} \tag{A9}$$

$$\frac{\mathrm{d}y}{\mathrm{d}t} = m_S \left(1 - \frac{y(t-5)}{N}\right)\left[y(t-5) - y(t-8)\right] \quad . \tag{A9b}$$

Equation (A9b) agrees exactly with Equation (29) of Ref. [9]; if we exclude the 5-day delay which merely shifts the infection curve rightwards by 5 days, it becomes the standard retarded logistic equation. Thus, the classical model emerges as a limiting case of the present model.

Although the DDE (A9b) is a special case of CST model and its predictive power is more limited, there are some occasions where (A9b) is more useful. For example, the averaging assumptions allow us to smoothly accommodate asymptomatic and symptomatic cases who have different transmissibility durations. This is harder to implement in CST model. Contact tracing can be easily incorporated into the DDE model [9] but it is more difficult here – a contact trace is expected to break transmission partway through a cluster; how can we account for partially infected clusters ? Cluster fragmentation is currently not incorporated into this model. Age-structuring is another phenomenon which is easier to model using differential equations. Finally, as mentioned in §14, rigorous lockdowns (full lock or at least closure of entertainment venues) are likely to drastically reduce the cluster size or make the cluster concept irrelevant altogether. In such a situation, the socially active people are expected to behave like millions of tiny clusters, which is better described by (A9b) than by the CST model.

§A4. Computer evaluation of (5), (6). While running the simulations, we observed that evaluation of the expressions (5) and (6) was impossible for the computer unless they were inputted in a special form. This happened because each of the probabilities $P(k)$, $P(b\,|\,k)$ and $P(j\,|\,b)$ features a ratio of two huge numbers, both of which are beyond the computer's calculational capacity. However, each huge number here is a product of many smaller numbers, and the computer has no trouble handling these individually. Thus, in the expression

$$P = \frac{n}{d} = \frac{n_1 n_2 .....n_m}{d_1 d_2 .....d_m} \quad , \tag{A10}$$

n and d are beyond the computer's resources while n_1, n_2, d_1, d_2 etc are not. To evaluate P, we must enter it as $(n_1/d_1) \times (n_2/d_2) \times .....\times (n_m/d_m)$ with n_i and d_i preferably being of comparable size, and then the computer has no difficulty in performing the calculation. Analytically opening out the combinations, we have written the probabilities *P*(*k*), *P*(*b* | *k*) and *P*(*j* | *b*) for easy machine evaluation in the forms described below.

The analytical formula for *P*(*k*) is

$$P(k)=\frac{{}^{\alpha}\mathrm{C}_{k}\ {}^{N_1-\alpha}\mathrm{C}_{n_S-k}}{{}^{N_1}\mathrm{C}_{n_S}} \quad , \tag{A11=1}$$

where we have repeated (1) of the main text as (A11) here just for display convenience. For the computer we define this as a function g = prob1(a,k,N,s) where N denotes N_1 and s denotes n_S. Opening out the combinations and rearranging items so that large numerators carry large denominators (note that N is the really large number here while the others are much smaller), we find

$$g=\underbrace{\left(\frac{N-a}{N}\right)\left(\frac{N-a-1}{N-1}\right).....\left(\frac{N-a-s+k+1}{N-s+k+1}\right)}_{s-k\text{ terms}}\times\underbrace{\left(\frac{a}{N-s+k}\right)\left(\frac{a-1}{N-s+k-1}\right).....\left(\frac{a-k+1}{N-s+1}\right)}_{k\text{ terms}}$$
$$\times\underbrace{\left(\frac{s}{k}\right)\left(\frac{s-1}{k-1}\right).....\left(\frac{s-k+1}{1}\right)}_{k\text{ terms}} \quad . \tag{A12}$$

Since k cannot be zero and is likely to be less than s (it is extremely improbable that every single person attending SEC gets infected), there are no borderline cases to be taken care of manually.

The analytical formula for *P*(*b* | *β*) is

$$P\left(b\,|\,\beta\right)=\frac{{}^{N_C}\mathrm{C}_{b}\ {}^{\beta-1}\mathrm{C}_{b-1}}{{}^{N_C-1+\beta}\mathrm{C}_{\beta}} \quad . \tag{A13=2}$$

For the computer we define the function g = prob2(N,b,t) where N denotes N_C and t denotes β. This function is

$$g=\underbrace{\left(\frac{N}{N+t-1}\right)\left(\frac{N-1}{N+t-2}\right).....\left(\frac{N-b+1}{N+t-b}\right)}_{b\text{ terms}}\times\underbrace{\left(\frac{t}{N+t-b-1}\right)\left(\frac{t-1}{N+t-b-2}\right).....\left(\frac{b+1}{N}\right)}_{t-b\text{ terms}}$$
$$\times\underbrace{\left(\frac{t-1}{b-1}\right)\left(\frac{t-2}{b-2}\right).....\left(\frac{t-b+1}{1}\right)}_{b-1\text{ terms}} \quad . \tag{A14}$$

Unlike with (A12), this has marginal cases. If t = 1 (impossible for SEC where *β* is defined as 2*k* but quite possible for UCT) then b must equal 1 and g must be unity. If b = t (a physically meaningful case) then the second product in (A5) is non-existent i.e. it must equal unity. If b = 1 (for whatever t) then the third product must be bypassed and set to unity.

The analytical formula for *P*(*j* | *b*) is

$$P\left(j\,|\,b\right)=\frac{{}^{N_C-z_i}\mathrm{C}_{j}\ {}^{z_i}\mathrm{C}_{b-j}}{{}^{N_C}\mathrm{C}_{b}} \quad . \tag{A15=3}$$

For the computer we define the function g = prob3(z,j,N,b) where z denotes z_i and N denotes N_C. We recognize that this is the same as the function g = prob1(z,b-j,N,b) except for the boundary cases which we recognize by looking at (A12). If b = j then g can be evaluated manually as

$$g=\underbrace{\left(\frac{N-z}{N}\right)\left(\frac{N-z-1}{N-1}\right).....\left(\frac{N-z-j+1}{N-z+1}\right)}_{j\text{ terms}} \quad . \tag{A16}$$

If z = N i.e. all clusters have been infected then g = 0. In all other circumstances, we define q = b-j and set g = prob1(z,q,N,b).